The Zinc Finger Associated Domain of *Drosophila melanogaster*, its Evolution and Phylogenetic Restriction

Von der Gemeinsamen Naturwissenschaftlichen Fakultät
der Technischen Universität Carolo-Wilhelmina
zu Braunschweig
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr.rer.nat.)
genehmigte
D i s s e r t a t i o n

von Ho Ryun Chung
aus Langen(Hessen)

1. Referentin oder Referent: Prof. Dr. Dieter Jahn
2. Referentin oder Referent: Prof. Dr. Herbert Jäckle
eingereicht am: 7.10.2004
mündliche Prüfung (Disputation) am: 17.12.2004

2005 (Druckjahr)

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Gemeinsamen Naturwissenschaftlichen Fakultät, vertreten durch den Mentor oder den Betreuer der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

Jauch, R., Bourenkov, G.P., **Chung, H.R.**, Urlaub, H., Reidt, U., Jäckle, H., and Wahl, M.C. (2003) The Zinc Finger-Associated Domain of the *Drosophila* Transcription Factor Grauzone is a Novel Zinc-Coordinating Protein-Protein Interaction Module. Structure **11**, 1393-1402

Chung, H.R., Schäfer, U., Jäckle, H., and Böhm, S. (2002) Genomic expansion and clustering of ZAD-containing C2H2 zinc-finger genes in *Drosophila*. EMBO Rep **3**, 1158-1162

Contents

List of Figures

List of Tables

Abbrevations

2L	left arm of the second chromosome
2R	right arm of the second chromosome
3L	left arm of the third chromosome
3R	right arm of the third chromosome
4	fourth chromosome
AgaG	*A. gambiae* genomic sequences
A. gambiae	*Anopheles gambiae*
Å	Ångstrom
Agam	*Anopheles gambiae*
BP	Bootstrap Proportion
BTB	*Broad-Complex*, *tramtrack*, *bric á brac*
CelDB	*Caenorhabditis elegans* protein database
CelHit	*Caenorhabditis elegans* hit to *D. melanogaster* ZFP
clustSeq	clustered sequence
CP	bootstrap Confidence Probability value for positive branch length
D. spec.	*Drosophila spec.*
Dana	*D. ananassae*
DbEST	EST databases
DIP1	*Dorsal Interacting Protein 1* gene
DIP1	Dorsal Interacting Protein 1 protein
Dmel	*D. melanogaster*
DmePR3.2	*D. melanogaster* proteome; release 3.2
DmePutZFP	putative *D. melanogaster* ZFPs
DmeZFP	*D. melanogaster* ZFPs
DM model	Dobzhansky-Muller model
DNA	DeoxyriboNucleic Acid
Dpse	*D. pseudoobscura*
DpeG	*D. pseudoobscura* genome sequences
dS	rate of synonymous exchanges per synonymous site
Dsim	*D. simulans*
Dvir	*D. virilis*
dwg	*deformed wings* gene
Dwg	Deformed wings protein
Dyak	*D. yakuba*
EST	Expressed Sequence Tag
e-value	expectation value indicating the probability of false-positive assignment
grau	*grauzone* gene
Grau	Grauzone protein
HC-link	interfinger spacer with the consensus sequence TGEKPF

HMM	(profile) Hidden Markov Model
HTH_7	Helix-turn-helix domain of resolvase (PF02796)
iZADMSA	initial ClustalW multiple sequence alignment
iZADHMM	initial ZAD HMM
Kr	*Krüppel* gene
Kr	Krüppel protein
KRAB	Krüppel associated box
MAPK	Mitogen Activated ProteinKinase
MmuDB	*Mus musculus* protein database
MmuHit	*Mus musculus* hit to *D. melanogaster* ZFP
mRNA	messenger-ribonucleic acid
N_{bp}	number of base pairs per genomic region
NFκB	Necrosis Factor κ B
NJ-tree	Neighbour joining tree
N_{ZAD}	number of ZAD-coding genes per chromosome or -arm
Pfam	database of Protein families
phyl	*phyllopod* gene
Phyl	Phyllopod protein
POZ	Poxvirus and zinc finger
PGC-2	PPAR γ (gamma) coactivator-2
PPAR γ	Peroxisome Proliferator-Activated Receptor γ
SCAN	SRE-ZBP, CTfin51, AW-1, Number 18 cDNA
Sry-δ	*Serendipity-δ* gene
Sry-δ	Serendipity-δ protein
SUBA*nn*	subgroup A ZAD of *D. pseudoobscura*, *nn* denotes the number 1-19
TFIIIA	(general) Transcription Factor III A
TraceDBDA	unassembled whole-genome shotgun sequence of *D. ananassae*
TraceDBDS	unassembled whole-genome shotgun sequence of *D. simulans*
TraceDBYA	unassembled whole-genome shotgun sequence of *D. yakuba*
TraceDBVI	unassembled whole-genome shotgun sequence of *D. virilis*
X	X chromosome
ZAD	Zinc finger Associated Domain
ZAD_{Grau}	ZAD of the transcription factor Grauzone
ZADHMM	final ZAD HMM
ZFP	C2H2 Zinc Finger Protein
ZFPg	genomic regions of *D. melanogaster* ZFPs

ZnF	C2H2 Zinc Finger

Pfam domains. In brackets Pfam accession number.

AT_hook	DNA binding motif with a preference for A/T rich regions (PF02178)
BAH	Bromo adjacent homology domain (PF01426)
BROMO	Bromodomain (PF00439)
BTB	*Broad-Complex*, *tramtrack*, *bric á brac* (PF00651)
CHROMO	CHRomatin Organisation MOdifier domain (PF00385)
DnaJ	DnaJ domain (PF00226)
DZF	found in proteins containing the double-stranded RNA-binding motif, DSRM or the zinc finger domain C2H2 (PF07528)
efhand	consists of an α-helix (E), loop, and second α-helix (F) (PF00036)
ELM2	Egl-27 and MTA1 homology 2 domain (PF01448)
FH	Fork head domain (PF00250)
FYRC	F/Y rich C-terminus (PF05965)
FYRN	F/Y rich N-terminus (PF05965)
GATA	GATA zinc finger (PF00320)
GoLoco	Galphai/o binding motif also found in the *D. melanogaster* protein encoded by locomotion defects (PF02188)
G-patch	has seven highly conserved glycines (PF01585)
HA2	Helicase associated domain 2 (PF04408)
Helicase_c	Helicase conserved C-terminal domain (PF00271)
HIT	Histidine Triad domain (PF01230)
HMG	high mobility group box (PF00505)
HOX	Homeobox domain (PF00046)
IPPT	IPP transferase (PF01715)
KOW	KOW motif, Kyprides, Ouzounis, Woese (PF00467)
LRR	Leucine rich repeat (PF00560)
LRV	Leucine rich repeat variant (PF01816)
MBD	Methyl-CpG binding domain (PF01429)
MOZ/SAS	monocytic leukemia Zinc finger protein or *Saccharomyces cerevisiae* protein involved in silencing (PF01853)
MYB	Myb-like DNA-binding domain (PF00249)
Otu	OTU-like cysteine protease (PF02338)
PHD	PHD-finger (PF00628)

PWWP	named after a conserved Pro-Trp-Trp-Pro motif (PF00855)
RHS	RHS protein (PF03527)
RRM_1	RNA recognition motif (PF00076)
SET	SET domain (PF00856)
Ssl1	Ssl1-like (PF04056)
THAP	THAP domain first identified in THAP1 (PF05485)
TNFR_c6	TNFR/NGFR cysteine-rich region (PF00020)
TPR	Tetratricopeptide repeat (PF00515)
Tubulin-binding	Tau and MAP protein, tubulin-binding repeat (PF00418)
TUDOR	Tudor domain (PF00567)
ubiquitin	Ubiquitin family (PF00240)
UBA	UBA/TS-N domain (PF00627)
UBX	UBX domain (PF00789)
UVR	UvrB/uvrC motif (PF02151)
zf-C2HC	zinc finger, C2HC type (PF01530)
zf-C3HC4	zinc finger, C3HC4 type also known as RING finger (PF00097)
zf-CCCH	zinc finger, C-x8-C-x5-C-x3-H type (PF00642)
zf-CCHC	Zinc knuckle (PF00098)
zf-RanBP	zinc finger in Ran binding proteins (PF00641)
zf-TRAF	zinc finger, TRAF-type (PF02176)
zf-U1	U1 zinc finger (PF06220)
ZZ	zinc finger, ZZ type (PF00569)

Chapter 1

Introduction

The completion of several metazoan whole genome sequencing projects has been one of the major scientific breakthroughs in the last decade (e.g. Adams et al., 2000; The *C. elegans* Sequencing Consortium, 1998; Lander et al., 2001). Their analysis, combined with the sequencing of Expressed Sequence Tags (ESTs) allowed scientists to obtain a nearly complete view on the proteomes of these species. Studies concerned with the molecular evolution of proteins and the underlying DNA sequences coding for them can now be conducted with an unprecedented wealth of data. These studies are an important contribution towards an understanding of the molecular basis of the evolutionary processes that led to the numerous species present today (Koonin et al., 2000). Phenotypic changes that accompany the evolution and divergence of species are due to changes in the genetic program encoded by the genome of an organism.

Transcriptional regulators are key components of the genetic program that directs the development of a fertilized egg to an adult organism (for an overview see Gilbert, 2003). Transcriptional regulators have therefore attracted and are still attracting considerable interest. The available whole genome sequences and the protein complement encoded by them allow comparisons of proteins involved in transcriptional regulation. For example, a comperative analysis of the transcriptional regulator families in three metazoan species, *Drosophila melanogaster* (*D. melanogaster*), *Caenorhabditis elegans* and *Homo sapiens*, revealed that only 30.5% of the transcriptional regulator families are present in all three species, 22.5% are shared by two species, whereas as much as 47%, i.e. nearly half of these families are species-specific (Coulson and Ouzounis, 2003).

Transcriptional regulators can be classified into two large classes: those that bind to DNA in a sequence-specific manner and others that act indirectly, e.g. by their association with a DNA binding protein. DNA binding domains, such as the homeodomain (McGinnis et al., 1984; Scott and Weiner, 1984), the basic helix-loop-helix domain (Murre et al., 1989) and the C2H2 zinc finger (ZnF) domain (Miller et al., 1985) can be used to further subgroup DNA-binding transcriptional regulators.

Here I present an *in silico* analysis of a group of transcriptional regulators of the zinc finger protein (ZFP) class initially identified in the *D. melanogaster* genome (Adams et al., 2000).

1.1 C2H2 zinc finger proteins

The ZnF motif was first identified in the *Xenopus laevis* basal transcription factor TFIIIA (Miller et al., 1985). The motif consists of 30 amino acids and is characterised by the presence of various combinations of four cysteine and/or histidine residues which mediate the coordination of a zinc ion (Berg and Shi, 1996; Klug and Schwabe, 1995). The coordination of a zinc ion allows the formation of an independently folding compact and stable 3 dimensional structure, which is composed of a β-hairpin followed by an α-helix (reviewed in Krishna et al., 2003). ZFPs have been, at large, implicated to mediate DNA binding in a sequence-specific manner (Pavletich and Pabo, 1991; Rosenberg et al., 1986; Wolfe et al., 2000). The ability to bind DNA in a sequence-specific manner is correlated to the presence of the so-called "HC-link", an interfinger spacer with the consensus sequence TGEKPF (Schuh et al., 1986).

Whole genome analysis has shown that ZFPs constitute the most abundant family of nucleic acid binding proteins in the eukaryotic kingdom (e.g. Lander et al., 2001). The functionally characterised members of the ZFP super family include proteins that act in the developing embryo of various organisms. One of the first examples of such a protein is encoded by the *D. melanogaster* segmentation gene *Krüppel* (*Kr*; Rosenberg et al., 1986). It belongs to the gap class of segmentation genes (Nüsslein-Volhard and Wieschaus, 1980; Pankratz et al., 1992) and is also necessary for the formation of various organs during embryogenesis (Hoch et al., 1990). Kr contains five ZnF domains that are connected by "HC-links" (see above) and has been shown to bind to DNA in a sequence-specific manner (Rosenberg et al., 1986; Schuh et al., 1986). Kr can be viewed as the prototype of ZnF-containing transcription factors (Schuh et al., 1986).

In a recent screen for putative *Krüppel*-dependent target genes using *in vivo* chromatin immunoprecipitation, a total of 83 candidate genes were identified (Matyash et al., 2004). The screen was far from saturation, since none of the known target genes have been identified by this method. It could be shown, however, that DNA-fragments corresponding to known Kr-dependent cis-regulatory elements were enriched in the Kr-dependent chromatin fraction. It has been subsequently approximated that *Kr* may regulate several hundred genes during the life-cycle of the fruitfly (Matyash et al., 2004). These results implied that ZFPs in general have the potential to regulate many target genes. Target specificity may be achieved by a certain combina-

tion of transcription factors assembled on so-called cis-regulatory sequences forming higher order complexes, called enhanceosomes (reviewed in Merika and Thanos, 2001), which in turn regulate transcription of adjacent genes.

The ability to form higher order complexes is critically dependent on interactions among the participating proteins. Many transcriptional regulators contain protein-protein interaction domains, like the BTB domain (*Broad-Complex*, *tramtrack*, *bric á brac*; Zollman et al., 1994) also known as POZ domain (Poxvirus and zinc finger; Bardwell and Treisman, 1994), the KRAB domain (Krüppel associated box; Bellefroid et al., 1991) and the SCAN domain (SRE-ZBP, CTfin51, AW-1, Number 18 cDNA; Williams et al., 1999; for a review of all three domains see Collins et al., 2001). These three domains are often associated with ZnF motifs and characterise subfamilies of ZFPs. Proteins of these subfamilies are often species-specific and represent a major fraction of the proteins of the ZFP class encoded by the genome of the respective organisms (Lander et al., 2001). For example the KRAB and SCAN domains are specific for vertebrates and proteins containing these domains are exceptionally abundant in the human genome (Lander et al., 2001). In particular, the KRAB-containing ZFPs have attracted considerable interest. Functional characterisation of the KRAB domain has shown that it functions as a transcriptional repressor domain by recruiting corepressor proteins (reviewed in Urrutia, 2003). The analysis of the evolutionary relationship of closely related KRAB-containing ZFPs in mice and humans showed that many human counterparts have single orthologs in mice. But it is often found that human KRAB-containing ZFPs have multiple orthologs in mice and vice versa (Mark et al., 1999; Shannon et al., 2003). The total lack of KRAB and SCAN domains in non-vertebrate species is remarkable, since the majority of domains seem to be shared by the taxa of the eukaryotic kingdom (Lander et al., 2001). The presence of these non-conserved domains in many proteins suggests that they have an adaptive advantage for the respective species.

1.2 Lineage-specific expansion of protein families

Comparative analysis of the proteomes of several representative species of the eukaryotic kingdom has revealed that members of some protein families have increased their number dramatically after the divergence of the major eukaryotic lineages. This phenomenon has been termed lineage-specific expansion (Lespinet et al., 2002). Lineage-specific expansions of protein families involve certain classes of proteins related to unique structural and biochemical properties of the respective lineage (e.g. cuticular proteins in insects and enzymes involved in the biosynthesis of pectin and cellulose in plants), but also other organism-specific functional classes, such as immune

response, transcriptional regulation, signal transduction and protein modification/degradation (Lespinet et al., 2002). It seems that many proteins involved in lineage-specific expansions have functions that are dependent on their ability to recognise distinct moieties, like molecules presented on the surface of pathogens, binding sites in DNA and target sites in proteins for posttranslational modifications. Most strikingly protein domains participating in lineage-specific expansions are often implicated to mediate protein-protein or nucleic acid-protein interactions, like the above mentioned KRAB, SCAN and ZnF domains.

Analysis of lineage-specific expansion has been performed with species that belong to taxa separated by large evolutionary distances (Lespinet et al., 2002). Similar analyses with species that are more closely related have been started, like a comparison of some KRAB-containing ZFPs in the mouse and human genome (Mark et al., 1999; Shannon et al., 2003), or a whole genome comparison of the two dipterans *Anopheles gambiae* (*A. gambiae*) and *D. melanogaster* (Zdobnov et al., 2002). A comparison of the copy number of 29,619 human genes across five hominid species (*Homo sapiens*, Human, *Pan troglodytes*, Chimpanzee, *Pan paniscus*, Bonobo, *Gorilla gorilla*, Gorilla, *Pongo pygmaeus*, Orangutan) revealed lineage-specific duplications and contractions of genes (Fortna et al., 2004). These result establish the idea that many genes are subject to lineage-specific expansion even in closely related species.

It remains a mystery why genes coding for these proteins exhibit such a dynamic evolutionary history, i.e. the expansion of families as well as the repeated loss and gain of distinct family members (e.g. Aravind et al., 2000; Lander et al., 2001; Shiu and Li, 2004). The majority of proteins that have undergone lineage-specific expansions contain domains that mediate interactions between biomolecules (see above). Genes coding for proteins of this class have been proposed to play a role in speciation, the so-called Dobzhansky-Muller model (DM model; Dobzhansky, 1936; Muller, 1939; Muller, 1940). A modern interpretation of the DM model predicts that genes that encode physically interacting proteins or proteins that assemble on cis-regulatory elements may be subject to divergent co-adaption, leading to the generation of incompatible alleles of interacting doublets, triplets or multiplets of higher order in nascent species. Incompatibilities are thought to be a by-product of the genetic divergence of isolated populations of a single species and contribute to the formation of reproductive barriers upon secondary contact. These mechanisms prevent gene flow between the nascent species, promote therefore the divergence of the genomes of these species and lead to the formation of new species. In this view, genes coding for members of lineage-specific expanded protein families might represent the raw material for the processes underlying the DM model.

Here I will present the identification of and an in-depth *in silico* analysis of the single largest subfamily of *D. melanogaster* ZFPs that is characterised by a distinct protein domain, termed zinc finger associated domain (ZAD; Chung et al., 2002). I will present evidence that the ZAD is a protein-protein interaction domain that is restricted to the insect lineage. I will discuss the mechanisms that led to the ZAD proteome of *D. melanogaster*. Moreover, the *in silico* comparison of the ZAD proteome encoded by the genomes of three dipteran flies, *D. melanogaster*, *Drosophila pseudoobscura* (*D. pseudoobscura*; http://www.hgsc.bcm.tmc.edu/projects/drosophila/) and *A. gambiae* (Holt et al., 2002) revealed that the genes coding for ZAD-containing proteins are subject to lineage-specific expansions. Finally, I will discuss evidence that might hint towards a role of ZAD-containing proteins in insect speciation.

Chapter 2

Material and Methods

2.1 Identification of C2H2 zinc finger proteins and zinc finger protein associated domains in the *D. melanogaster* genome

In order to identify ZFPs in the *D. melanogaster* proteome, the Pfam domain PF00096 (Bateman et al., 2004) in conjunction with `hmmsearch` of the HMMER package (version 2.1.1; Eddy, 1998) was used to search the *D. melanogaster* proteome (Release 3.2; DmePR3.2). A score 0.0 was chosen as a threshold:

```
>hmmsearch PF00096 DmePR3.2.
```

The resulting set of putative ZFPs (DmePutZFP) was extracted (perl script `parseHMMER.pl` on the CD) and subsequently used to search against the whole Pfam A database (release 14; Bateman et al., 2004) using `hmmpfam` of the HMMER package:

```
>hmmpfam PfamA DmePutZFP.
```

The identified ZnF motifs were subsequently checked for overlaps to other protein motifs and/or domains. Putative C2H2 zinc finger motifs that overlap more significant hits to other protein motifs were eliminated.

The resulting set of proteins was checked for identical splice-variants. In case of identical protein sequences only one was kept. This process yielded a total number of 454 non-redundant ZFPs in the *D. melanogaster* proteome encoded by 359 genes. Hits to additional domains in the Pfam A database were extracted and annotated for each protein if their e-value was less than 1.0 (perl script `parsePfam.pl` on CD).

2.2 Multiple sequence alignments and profile construction

An initial set of ZADs was used to generate a multiple sequence alignment (iZADMSA) with ClustalW (version 1.81; Thompson et al., 1994). This multiple sequence alignment was used to construct a profile hidden Markov model (iZADHMM) using `hmmbuild` and `hmmcalibrate` of the HMMER package with default settings:

`>hmmbuild iZADHMM iZADMSA`
`>hmmcalibrate iZADHMM`.
A search against the individual genomic regions (ZFPg) of the identified ZFPs was performed using `genewise` of the Wise package (version 2.2.0; Birney et al., 1996):
`>genewise -hmmer iZADHMM ZFPg`.
The genomic structure of the identified ZAD-containing ZFPs was determined (if possible) using the Gene2EST package (Gemünd et al., 2001) in combination with BLAST (version 2.2.0; Altschul et al., 1997). The verified ZAD-coding protein sequences were aligned using ClustalW with default parameters. This multiple sequence alignment was used to construct an enhanced profile HMM (ZADHMM; see above).

2.3 Identification and annotation of ZAD-coding sequences in the genomes of *D. pseudoobscura* and *A. gambiae*

The ZADHMM was used to performed searches against the genomic sequences of *D. pseudoobscura* (DpeG; freeze assembly 02.27.2003; http://www.hgsc.bcm.tmc.edu/projects/drosophila/):
`>genewisedb -hmmer ZADHMM DpeG`,
and *A. gambiae* (AgaG; Holt et al., 2002):
`>genewisedb -hmmer ZADHMM AgaG`,

Both, the *D. pseudoobscura* and the *A. gambiae* genome sequences were fractionated into 2,000 base pair fragments which overlapped by 200 base pairs (perl script `fractionate.pl` on the CD) and divided into files with 1,000 sequences each (perl script `divide.pl` on the CD).

For all putative genomic sequence intervals of *D. pseudoobscura* and *A. gambiae* that had a positve `genewise` hit, gene predictions using `GenScan` (Burge and Karlin, 1997) were performend (perl script `RunGenScan.pl` on the CD). In addition `genewise` was used to find orthologs of *D. melanogaster* ZADs in the *D. pseudoobscura* genome. The annotations produced by these programs were manually inspected to delete annotation errors using Artemis (Rutherford et al., 2000).

2.4 Searches for ZAD-coding sequences in EST databases

The enhanced ZAD HMM was used to search in EST databases for ZAD-coding sequences. This was done using `estwisedb` of the Wise package:
`>estwisedb -hmmer ZADHMM ESTDB`.
For each hit in the DbEST databases (see Table 2.1) the GI number was extracted (perl script `ExtractGI.pl` on the CD). The list of GI numbers

was used to download the EST sequences in GenBank format from GenBank using the batch entrez mode. The species name was extracted from these files (perl script `GetOrganism.pl` on the CD). The screened EST databases are sumarised in Table 2.1.

Table 2.1: Screened EST Databases and their origin.

Database	origin	download time
DbEST	GenBank http://www.ncbi.nlm.nih.gov/	May 2002
DbEST Arthropoda	GenBank http://www.ncbi.nlm.nih.gov/	February 2004
ESTs *Daphnia spec.*	ftp://iubio.bio.indiana.edu/daphnia/	February 2004

2.5 Identification of ZAD-coding sequences of subgroup A in whole-genome shotgun sequences of four other Drosophilid species

In order to date the ZAD-coding genes of subgroup A, which have been shown to be differentially expanded in *D. melanogaster* and *D. pseudoobscura*, I have searched orthologous ZAD-coding sequences in the unassembled whole-genome shotgun sequences of *D. simulans* (TraceDBDS), *D. yakuba* (TraceDBDY), *D. ananassae* (TraceDBDA) and *D. virilis* (TraceDBDV). The trace sequences have been downloaded from ftp://ftp.ensembl.org/traces/ in Mai 2004. All ZAD-coding sequences of subgroup A of *D. melanogaster* and *D. pseudoobscura* (ZADSubA) were used as query:
`>blastall -p tblastn -i ZADSubA -d TraceDB(DS, DY, DA or DV)`.
The identified trace sequences (trace_i) were used as input to genewise in order to extract the ZAD-coding sequences:
`>genewise -hmmer ZADHMM` trace_i
where the subscript i denotes a single trace sequence (perl script `trace2ZAD.pl` on the CD). All identified ZAD-coding sequences were clustered by
`>blastclust -i identified ZAD-coding sequences`.
The clustered sequences (ClustSeq) were aligned and a consensus sequence was determined (perl script `getbestZAD.pl` on CD). All identified protein sequences were compared to all *D. melanogaster* and *D. pseudoobscura* ZADs of subgroup A and homologous sequences were extracted and used in the Neighbour-joining tree construction.

2.6 Tree construction and statistical analysis

Starting from a seed alignment (Figure 3.1 A) all other ZAD sequences were added by the profile alignment mode of ClustalW using default parameters. Tree construction and statistical analysis was performed using MEGA (version 2.1; Kumar et al., 2001). All trees presented in this work are neighbour joining trees and the reliability of groups was tested using the bootstrap and the interior branch test (within MEGA) with 500 replicas each.

2.7 Identification of orthologs in *Mus musculus*, *Caenorhabditis elegans* and *A. gambiae*

Each *D. melanogaster* ZFP (DmeZFP) was compared against all *Mus musculus* (MmuDB) and *Caenorhabitis elegans* (CelDB) proteins (downloaded February 2004 from http://www.ncbi.nlm.nih.gov/ using the query strings "Mus musculus[ORGN]" and "Caenorhabditis elegans[ORGN])" in the GenBank protein database). This was done using the `blastall` program of the BLAST package using the *D. melanogaster* ZFPs as query:
`blastall -p blastp -i DmeZFP -d MmuDB` or
`blastall -p blastp -i DmeZFP -d CelDB`.
The resulting best match in the *Mus musculus* (MmuHit) and *Caenorhabditis elegans* (CelHit) protein database was compared to all *D. melanogaster* proteins:
`blastall -p blastp -i MmuHit -d DmePR3.2` or
`blastall -p blastp -i CelHit -d DmePR3.2` .
The queried *D. melanogaster* ZFP was assigned to have an ortholog in *Mus musculus* and/or *Caenorhabditis elegans* if the best hit of the second BLAST search was again the queried sequence (perl script `Orthology.pl` on the CD).

The same procedure was used to approximate the number of orthologs of ZAD-containing proteins in *D. melanogaster* and *A. gambiae* with an additional constraint that the two hits had to be at least 30% identical and the match length was longer than 100 amino acids. This is essentially the same method used in Zdobnov et al. (2002) and allows thereby the comparison of their data to the data obtained here.

2.8 Calculation of similarities of the ZAD and the remaining protein sequence of subgroup A

The ZAD amino acid sequences of subgroup A ZADs of *D. melanogaster* and *D. pseudoobscura*, respectively, were aligned using Dialign2 (Morgenstern, 1999; http://www.genomatix.de/cgi-bin/dialign/dialign.pl) with the "Additional output of pairwise sequence similarities" option. The same procedure was used for the remaining portion of the proteins. The similarity tables

were used to calculate the average similarity and the standard deviation of the ZAD amino acid sequence and the remaining protion of the proteins. The same calculation was repeated using only the most similar pairs of sequences for each gene.

2.9 Identification of clustered ZAD-coding genes in the *D. melanogaster* genome

The positions of ZAD-coding genes in the *D. melanogaster* genome sequence Release 3 (Celniker et al., 2002) were extracted from Release 3.2 annotations. Given the unequal distribution of ZAD-coding genes on the chromosomes and -arms, the test was conducted for the X, 2L, 2R, 3L, 3R contigs separately.

The total number N_{ZAD} of ZADs per chromosome (X chromosome) and chromosome-arm (left and right arm of the second and third chromosome) was determined and divided by the number of base pairs N_{bp} of corresponding chromosome or chromosome-arm (see Celniker et al., 2002 for the sizes of the chromosomal regions). This frequency, $p = \frac{N_{ZAD}}{N_{bp}}$, was used as the rate ZAD/base pair of a exponential distribution. The number of ZADs in a given sequence interval of N base pairs would follow a Poisson distribution, if the number of ZADs in a given sequence intervall is completely random, one would therefore expect to find $\langle ZAD \rangle = Np$. $\langle ZAD \rangle$ was compared to the observed numbers. The squared difference between expected and observed numbers was divided by the expected number and gave rise to a value that follows the χ^2 distribution with one degree of freedom. The set of genes that maximised this value and gave a p-value $<$ 1e-10 was assigned to be significantly clustered. In one case I included a fourth gene although the χ^2 value was lower, but on visual inspection this gene clearly belongs to this clusters (this cluster is marked with † in Table 3.2).

2.10 Identification of orthologs of *D. melanogaster* genes in *D. pseudoobscura*

For each annotated gene of *D. melanogaster* (Release 3.1) the longest protein sequence was chosen. This set of protein sequences was used to search for homologous sequences in the *D. pseudoobscura* genome using the tblastn mode of Blast:

```
>blastall -p tblastn -i DrosophilaMelanogasterProteins
-d DrosophilaPseudoobscuraGenome.
```

The best matching contig was extracted from the whole genome sequence database if the e-value was lower than 1e-10 and a subsequence of ± 10,000 base pairs of the start and end of the hit were taken. These subsequences were used to screen with `genewise` of the Wise package using the *D. me-*

lanogaster protein sequence as query. The predicted protein sequence and the predicted coding sequence (nucleotide sequence) were extracted from the `genewise` output (perl script `dme2genome.pl` on the CD). All orthologous pairs that involved the same genomic sequence of *D. pseudoobscura* and two or more protein sequences of *D. melanogaster* were further analysed using the approach of reciprocal blasts (see Materials and Methods 2.7) to delete redundant hits. This approach gave rise to 11,099 orthologous pairs.

2.11 Determination of the rate of synonymous exchanges dS

The protein sequences of all ortholog pairs were aligned using ClustalW. This amino acid alignment was mapped to the DNA coding-sequences (perl script `formatAlignment.pl` on the CD). The rate of synonymous exchanges dS of all ortholog pairs was determined using `codeml` (seqtype 1, codons; runmode -2, pairwise comparison; CodonFreq 2, F3X4 see Yang (1997) for details) of the PAML package(version 3.2; Yang, 1997). In a second run the codon frequencies were held constant (same parameters except CodonFreq 0, 1/61; perl script `runPAML.pl` on the CD). All PAML dS values were extracted from the output of codeml as well as the dS values calculated after Nei and Gojobori (1986) (perl script `ParsePAML.pl` on the CD). There were a total of 11,099 gene pairs for the PAML dS value dataset and 10,414 of the dS values computed after Nei and Gojobori (1986). The discrapancy between these two numbers is due to some sequence pairs that were too diverged to be analysed using the method of Nei and Gojobori (1986). The data was bined into 500 bins from dS values of 0 to 102 and a normalized histogram was determined using xmgrace (http://plasma-gate.weizmann.ac.il/Grace/).

2.12 Secondary structure prediction

Secondary structures were predicted using PHD (Rost, 1996). The multiple sequence alignment shown in Figure B.1 was used as input to PHD. For each ZAD-coding sequence the secondary structure was predicted by puting each sequence at the top position of the alignment (perl script `runPHD.pl` on the CD). The output was remapped to the sequence alignment and the "consensus secondary structure prediction" was calculated using the PHD scores as weights (perl script `ConsensusPHD.pl` on the CD).

2.13 Estimating the divergence time for the Drosophila species

In order to calculate the divergence time for the 6 Drosophila species, I repeated the analysis of Russo et al. (1995). The analysis was performed with

alcohol dehydrogenase genes available from GenBank (see Russo et al., 1995 for accession numbers) and an additional *alcohol dehydrogenase* sequence of *D. ananassae* extracted from the unassembled whole-genome shotgun sequences:

D. ananassae Alcohol dehydrogenase N-terminal fragment
Protein sequence

```
MALSLTNKNVVFVAGLGGIGLDTTKELLKRDLKNLVILDRIDNPVVIAELKTINPKVTVT
FIPYDVTVPITETKKLLKTIFDKLKTVDILINGAGILDDHQIERTIAVNYTGLVNTTTAI
LDFWDKRKGGPGGIICNIGSVTGFNAIYQVPVYSGTKAAVVNFTSSLAKLAPITGVTAYT
VNPGITRTTLVHKFNSWLDVEPTQSSQACGENFVKAIELNQNGAIWKLDRGTLEAIQWSK
HWDSGI
```

Coding sequence

```
ATGGCACTATCACTCACCAACAAGAACGTGGTCTTCGTGGCTGGCCTGGGAGGCATTGGC
CTGGACACCACCAAGGAGCTGCTCAAGCGCGACCTGAAGAACCTGGTGATCCTGGACCGC
ATCGACAACCCGGTTGTCATTGCCGAGCTGAAGACAATCAATCCCAAGGTGACCGTCACC
TTCATTCCCTACGATGTGACCGTGCCCATTACCGAGACCAAGAAGCTGCTGAAGACCATC
TTCGACAAGCTGAAGACCGTGGACATCCTGATCAACGGAGCTGGCATCCTGGATGACCAC
CAGATCGAGCGCACCATCGCCGTCAACTACACGGGTCTGGTGAACACCACCACCGCCATT
CTGGACTTCTGGGACAAGCGCAAGGGCGGACCAGGTGGCATCATCTGCAACATTGGCTCC
GTAACCGGCTTCAACGCCATCTACCAGGTGCCCGTCTACTCTGGCACCAAGGCTGCCGTT
GTCAACTTCACCAGCTCCCTAGCCAAACTGGCTCCCATCACCGGAGTGACCGCCTACACC
GTGAACCCCGGCATCACCCGCACCACCCTGGTGCACAAGTTCAACTCCTGGCTGGATGTG
GAACCCACCCAGTCCTCCCAGGCCTGCGGCGAGAACTTCGTCAAGGCCATCGAGCTCAAC
CAGAACGGCGCCATCTGGAAGCTCGACCGGGGCACCCTGGAAGCCATCCAGTGGAGCAAG
CACTGGGACTCCGGCATC
```

A neighbour-joining tree was calculated using MEGA (see above). The third position of each codon was used to calculate the distance matrix with the distance metric of Tajima and Nei (see Russo et al., 1995 for details). The divergence time was then found by setting the divergence time of *D. picticornis* and *D. silvestris* to 5.1 million years (see Russo et al. (1995) for details).

2.14 Hardware and additional Software

Most programs were run on an AMD Athlon(tm) XP 2000+ Linux machine running SuSe Linux (version 9.0; SUSE LINUX AG, Nürnberg, Germany). The exceptions are, the MEGA2 program was run on an AMD Athlon(tm) 1000 Windows machine running Microsoft Windows XP Professional 2002 (Service Pack 1; Microsoft Corp., Redmond, WA 98052-6399, USA), the Pfam searches were run on a Linux cluster with 54 nodes (Dual Intel Xeon(tm)

Table 2.2: Estimated divergence times between *D. melanogaster* and the other studied Drosophilids.

Taxa compared	Time [million years]
D. melanogaster vs *D. simulans*	2.7
D. melanogaster vs *D. yakuba*	7.3
D. melanogaster vs *D. ananassae*	16.7
D. melanogaster vs *D. pseudoobscura*	30.4
D. melanogaster vs *D. virilis*	47.1

3,06 GHz) at the Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen in Göttingen, Germany.

The figures were prepared with Macromedia Freehand (version 8.01; Macromedia Inc., San Francisco, CA 94103, USA) on an Apple Power Mac G4 running MacOS 9.2 (Apple, Cupertino, CA 95014, USA). The cartoon representation in Figure 3.1 B of the structure of the ZAD of Grauzone was prepared with RasMol (version 2.7.1; Sayle and Milner-White, 1995), MolScript (version 2.1.2; Kraulis, 1991) and Raster3D (version 2.7b; Merritt and Murphy, 1994). The surface representations in Figure 3.1 C-D were prepared using VMD (version 1.8.1; Humphrey et al., 1996) and Raster3D. The overlay in Figure 3.1 D was prepared using Adobe Photoshop (version 5.0.2; Adobe Systems Inc., San Jose, CA 95110-2704, USA) on an Apple Power Mac G4 running MacOS 9.2. The NJ trees were converted from MEGA format to Newick format by the MEGA3 program (Kumar et al., 2004). All NJ trees were visualised using TreeView (version 1.6.6; Page, 1996) and imported to Macromedia Freehand using PrintToPDF (http://www.jwwalker.com/pages/pdf.html).

Some of the perl scripts to perform the analyses used some functionality provided by BioPerl (version 1.4; http://www.bioperl.org). All scripts can be found along with their documentation on the accompanying CD.

Chapter 3

Results

3.1 Characterisation of C2H2 zinc finger protein-coding genes in the *D. melanogaster* genome

The *D. melanogaster* genome has been sequenced and presented in April 2000 (Release 1; Adams et al., 2000). The most recent sequence release (Release 3; Celniker et al., 2002) and new annotations (Release 3.2) include annotations for 13,543 protein coding genes and 18,746 proteins. The difference in the number of genes and the number of proteins is due to alternative splicing and alternative promotor usage, leading to different mRNAs and in turn to different proteins.

I found that the *D. melanogaster* genome contains 359 genes coding for a total of 454 ZFPs, i.e. extrapolated to the whole genome 2.6% of the genes and 2.4% of the proteins (see Table A.1 for details). The majority (251 of 454 = 55%) of these ZFPs are annotated as transcriptional regulators by Flybase (http://www.flybase.net), including functionally characterised transcription factors, such as *Krüppel* (Rosenberg et al., 1986), chromatin modifying enzymes, such as *Males absent on first* (Hilfiker et al., 1997) and proteins involved in the formation of so-called insulator complexes, such as *Deformed wings* (Fahmy and Fahmy, 1959; Gaszner et al., 1999).

Of the 359 ZFP-coding genes 114 genes coding for 144 proteins (31.7% of all *D. melanogaster* ZFP-coding genes) are conserved in *Mus musculus* and/or *Caenorhabditis elegans*, representing the vertebrate and nematode lineages (see Table A.1). The remaining 245 genes code for 310 ZFPs (68.3% of all *D. melanogaster* ZFP-coding genes) and seem to be specific for the arthropod crown group. These lineage-specific ZFPs can be grouped by the presence of additional protein motifs associated with the ZnF motifs, including two major groups of *D. melanogaster*-specific ZFPs (see Table A.1 for a detailed overview). One group composed of 15 genes (= 50 proteins) contain the evolutionary conserved BTB/POZ domain, a motif which has been implicated to mediate homodimerisation as well as in some instances heteromeric association of proteins (Huynh and Bardwell, 1998). The second group contains 94 genes (= 95 proteins) characterised by the presence of a N-terminal motif of 71-97 amino acids. In all but four cases, this motif can be found

associated with ZnF domains. Therefore, and albeit the notion that the domain has been recently termed C4DM domain (Lander et al., 2001; Lespinet et al., 2002), the motif will be referred to as zinc finger associated domain (ZAD). With 94 out of 359 genes (26.1%), the ZAD-containing ZFPs represent the single largest subgroup of ZFP-coding genes in the *D. melanogaster* genome.

3.2 The zinc finger associated domain

Despite the large number of 98 ZAD-coding genes in the *D. melanogaster* genome (94 ZADs associated with ZnF domains and four, for which no associated ZnF motif could be identified in the annotated protein sequence; see Table A2), only five of them have been identified by means of mutant alleles. Four of those genes, *deformed wings* (*dwg*; Fahmy and Fahmy, 1959), *grau-zone* (*grau*; Schüpbach and Wieschaus, 1989), *Serendipity-δ* (*Sry-δ*; Payre et al., 1990) and *pita* (*pita*; Laundrie et al., 2003), contain in addition to the ZAD ZnF domains, whereas the fifth gene, *phyllopod* (*phyl*; Chang et al., 1995; Dickson et al., 1995) has no other domains apart from a ZAD-like domain.

The functional characterisation and the results of biochemical studies on the first three genes suggest that they are involved in transcriptional regulation. *dwg* encodes a sequence-specific DNA-binding factor that promotes the formation of so-called insulator complexes (Gaszner et al., 1999). *grau* and *Sry-δ* encode transcription factors implicated in the activation of *cortex* (Chen et al., 2000; Harms et al., 2000) and *bicoid* (Payre et al., 1994), respectively. *pita* has been shown to be required for proper oogenesis and larval survival (Laundrie et al., 2003). Direct evidence for an involvement of Pita in transcriptional regulation is currently not available. Finally, the protein encoded by *phyl* has been suggested to be involved in the *seven in absentia*-dependent degradation of the transcriptional repressor encoded by the gene *tramtrack* by the ubiquitin-dependent protein degradation pathway (Li et al., 1997; Tang et al., 1997). It has been shown that *phyl* is transcriptionally upregulated in response to the Raf/MAPK (Chang et al., 1995; Dickson et al., 1995) signalling pathway during the determination of the R7 photoreceptor cell. Furthermore, it has been suggested that *phyl* is downregulated in response to Notch signalling during sensory organ precursor development (Pi et al., 2001) by Notch-dependent downregulation of the proneural genes *achaete* and *scute* which are necessary for *phyl* expression (Pi et al., 2004). Two additional ZAD-coding genes, termed *Dorsal interacting protein* (*DIP1*; Bhaskar et al., 2000) and *Neu2* (Stathopoulos et al., 2002), have been isolated. DIP1 has been shown to bind to the NFκB homolog *Dorsal* (Bhaskar et al., 2000) and *Neu2* has been proposed to be a target gene of *Dorsal*

(Stathopoulos et al., 2002). Collectively, the experimental evidence favours the proposal that ZAD-containing proteins are directly or indirectly (Phyl) involved in transcriptional regulation.

3.2.1 Properties of the ZAD

ZADs vary in length between 71 and 97 amino acid residues. A multiple sequence alignment of a representative subset of 18 ZADs (Figure 3.1 A) shows that the motif is composed of four conserved sequence blocks (block 1-4). They are linked by three variable regions (r1-r3) of variable length. The most striking feature of the ZAD is the occurrence of two invariant cysteine pairs in the blocks 1 and 4. Their strict conservation suggests that they coordinate the binding of a zinc ion to stabilise a distinct fold of the domain represented by the ZAD sequence motif. Furthermore, secondary structure analysis of all ZADs of *D. melanogaster* predicts that the variable regions 1-3, which contain preferentially small and polar amino acids (Figure 3.1 A), represent turns or unstructured spacers, whereas the conserved blocks 1-4 form a $\beta 1 \alpha 1 \alpha 2 \beta 2 \alpha 3$-fold (Figure 3.1 A). Within each of the blocks 1-4, most conserved amino acid residues are hydrophobic; the few exceptions include a highly conserved arginine residue (position 4; Figure 3.1 A) which is located between the first invariant cysteine pair of block 1.

The biological significance of the ZAD and of the arginine residue in particular is exemplified by the finding that a point mutation of *dwg*, which results in an arginine-to-glycine replacement, causes a lethal phenotype (Gaszner et al., 1999). Furthermore, a point mutation in *Sry-δ* that leads to a tyrosine replacement of the second invariant cysteine of block 1 causes a lethal phenotype as well (Crozatier et al., 1992). These observations suggest that the core structure of the ZAD carries an essential function at least in the case of *Sry-δ* and *dwg*. Mutational analysis combined with biochemical studies showed that the ZAD of Sry-δ functions as protein-protein interaction domain (Payre et al., 1997), a function that has been proposed for the ZAD of Dwg as well (Gaszner et al., 1999). Collectively, these observation led to the proposal that (i) the ZAD motif represents an independently folding domain, (ii) this domain is stabilised by the coordination of a zinc ion by the four invariant cysteines, and (iii) ZAD is a protein-protein interaction module. In order to test this proposal we determined the 3 dimensional structure of the ZAD of the transcription factor Grauzone (ZAD_{Grau}, amino acids 2 to 81) by X-ray crystallography at 2.0 Å resolution. In collaboration with Ralf Jauch, we were able to show that ZAD_{Grau} is a zinc-binding protein module and that zinc-binding significantly contributes to the stability of the fold (Jauch et al., 2003).

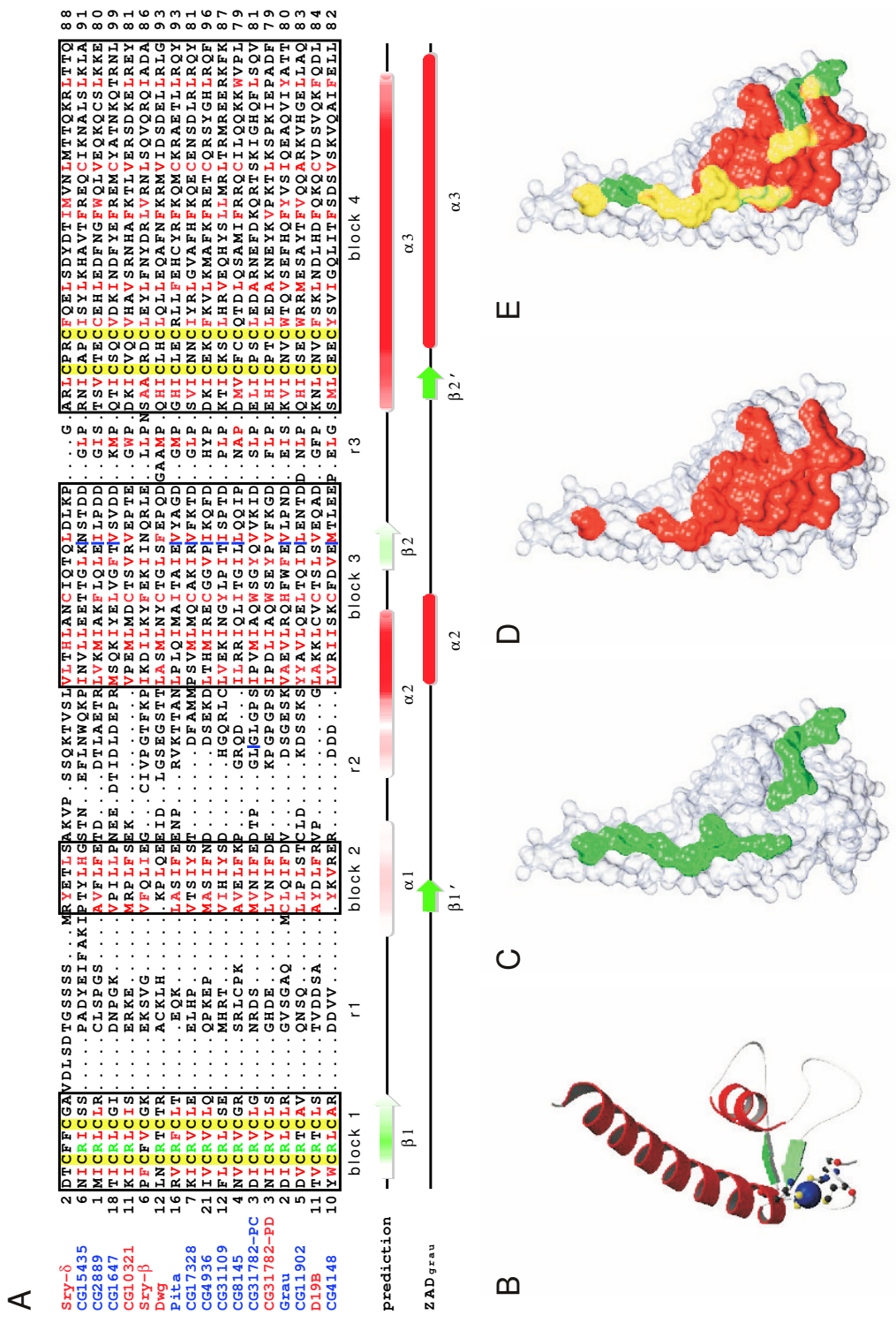


Figure 3.1 *legend on next page*

Figure 3.1: Properties of the ZAD. (A) Alignment of representative ZADs and comparision between secondary structure prediction and secondary structure elements of the ZAD_{Grau} structure; gene names in red indicate single exon ZADs, gene names in blue indicate two exon ZADs; yellow boxes, invariant cysteine residues; green characters, highly conserved arginine residues; red characters, conserved (>60%) hydrophobic amino acid residues; blue vertical lines indicate the position of the intron; conserved blocks 1-4 are framed; r1-r3 denote the variable regions 1-3; green arrows β-strands; red cylinders α-helices; significance of the secondary structure prediction is indicated: the darker green (red) the better the prediction. (B) cartoon of the ZAD_{Grau} structure. Colour-scheme as in A; the zinc ion is depicted as blue ball; the N-terminus is located at the bottom, the C-terminus at the top. (C)-(E) comparison of the amino acids constituting the contact interface and the conserved hydrophobic amino acids. (C) amino acids constituing the contact interface (Jauch et al., 2003) in green. (D) conserved hydrophobic amino acid residues in red. (E) merged image, showing the overlap between C and D in yellow.

The 3 dimensional structure (Figure 3.1 B) of ZAD_{Grau} represents a novel fold that is structurally related to the so-called treble-clef zinc finger fold (Jauch et al., 2003). Remarkably, the amino acids, which could be solved by X-ray crystallography, correspond to the amino acids included in the multiple sequence alignment shown in Figure 3.1 A. Amino acid 81 of ZAD_{Grau} already corresponds to the linker region that connects the ZAD with the rest of the protein (Jauch et al., 2003). This finding indicates that the boundaries set by the multiple sequence alignment shown in Figure 3.1 A correspond to the actual borders of the domain. In contrast to the predicted "consensus secondary structure elements", ZAD_{Grau} forms a $\beta1'\alpha2\beta2'\alpha3$-fold.

In Figure 3.1 A, a comparison between the predicted "consensus secondary structure elements" and the secondary structure elements determined by the crystal structure of ZAD_{Grau} (Jauch et al., 2003) is shown. The two predicted β-sheets, $\beta1$ and $\beta2$, were predicted with low confidence. These weak predictions could not be confirmed by the crystal structure of ZAD_{Grau}. The predicted α-helix $\alpha1$ is also weakly predicted and the crystal structure revealed a β-sheet ($\beta1'$) at the corresponding position. The remaining two α-helices, $\alpha2$ and $\alpha3$, are strongly predicted. The most significant portions of the predicted α-helices overlap with the helices determined by the crystal structure of ZAD_{Grau}. The N-terminal amino acids of the predicted α-helix $\alpha3$, however, form a β-sheet ($\beta2'$) in the structure of ZAD_{Grau}. The secondary structure predictions suggested that the conserved amino acids in blocks 1-4 form secondary structure elements, while the variable linker regions r1-r3 form unstructured spacers or represent turns. These results could be confirmed by the 3 dimensional structure of ZAD_{Grau}. The secondary structure elements of ZAD_{Grau} reside within the conserved sequence blocks, while the variable regions correspond to unstructured spacers or turns. Given the conservation of length, the location of secondary structure elements in conserved

sequence blocks (see Figure 3.1 A) and the conservation of critical amino acid residues of ZADs, the crystal structure of ZAD_{Grau} might therefore represent the prototype of the ZAD-fold (Jauch et al., 2003).

The proposal that the ZAD represents a protein-protein interaction domain, is supported by the finding that two ZAD_{Grau} molecules form a dimer in the crystal state. The contact between the two subunits buries around 20% of the accessible surface area (Jauch et al., 2003). Significantly, a large number of amino acid residues responsible for the contacts between the two subunits are identical to the aforementioned conserved hydrophobic amino acid residues (see Figure 3.1 C-E). Furthermore, biochemical and biophysical studies showed that ZAD_{Grau} also forms homodimers in solution (Jauch et al., 2003).

Taken together, the data derived by the determination of the 3 dimensional structure of ZAD_{Grau} as well as biochemical and biophysical approaches support the proposal that ZAD_{Grau} and possibly all other ZADs are independently folding domains that are stabilised by zinc-coordination. They most likely mediate homodimer formation and/or heterodimer association of closely related members of the ZAD family.

3.2.2 The ZAD is insect-specific

The ZAD characterises a family of ZFP genes that are not conserved in the genomes of the yeast *Saccharomyces cervisea* (Goffeau et al., 1996), the plant *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), the nematode *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium, 1998) and the vertebrate *Homo sapiens* (Lander et al., 2001; Venter et al., 2001). This initial finding suggests that ZADs are specific for the arthropod lineage. In order to test this hypothesis, I preformed a search for the domain in the publicly available ESTs provided by GenBank (http://www.ncbi.nlm.nih.gov/) and alternative sources (for details see Materials and Methods) using the program `estwisedb` of the Wise package (Birney et al., 1996; see Materials and Methods).

The results of this search were unambiguous. They show that ZAD-coding sequences can only be found in the ESTs of the insect orders *Diptera*, *Lepidoptera*, *Hymenoptera* and *Hemiptera* (Figure 3.2). Significantly, I was not able to find a single ZAD-coding sequence in over 7 million vertebrate ESTs and over 500,000 nematode ESTs (Figure 3.2). This suggests that these lineages do no have this domain or even distantly related sequences. Moreover, there was no ZAD-coding sequence detectable in the sister groups of the insect lineage, where >20,000 ESTs of *Crustacea* and >10,000 ESTs of *Chelicerata* were analysed (Figure 3.2). The lack of ZAD-coding sequences in the transcriptom of these non-insect species among the arthropods seems to

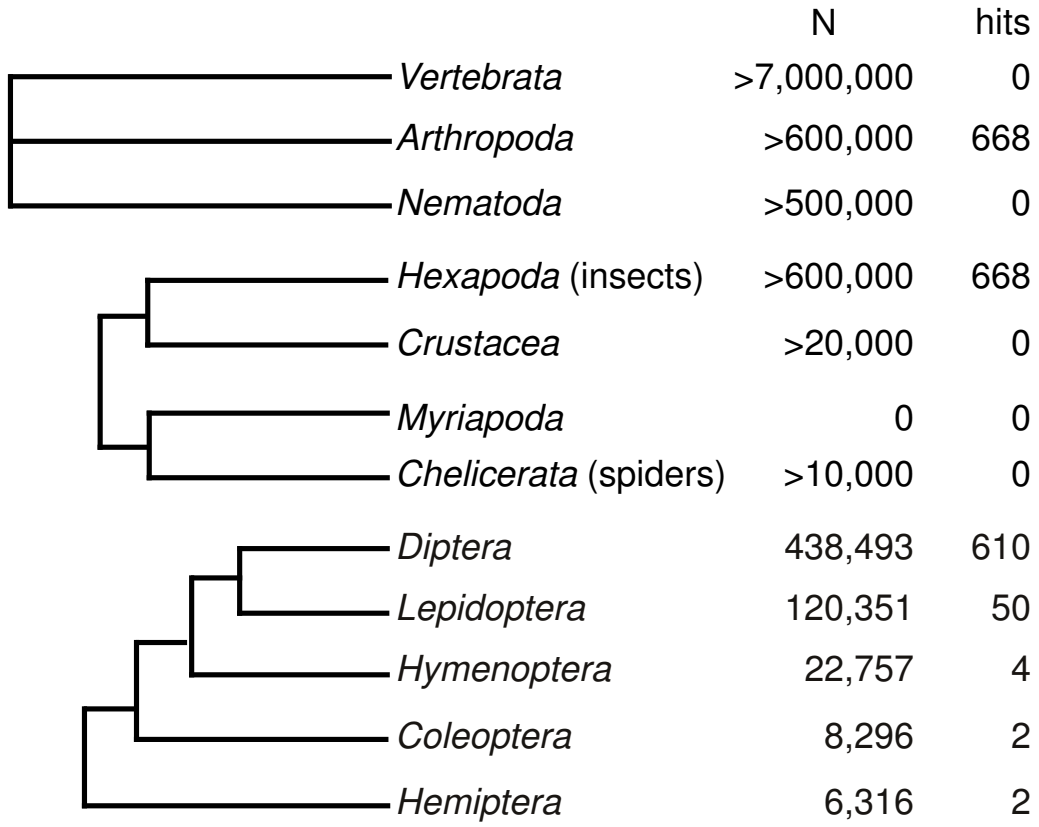


Figure 3.2: Phylogenetic distribution of the ZAD. N indicates the number of ESTs screened, and hits the number of positive identifications of ZADs.

be significant since a search in only 1,500 ESTs of the mosquito *Aedes aegyptii* yielded already 4 hits. However, the presence of ZAD-like sequences in the genomes of arthropod but non-insect species cannot be ruled out definitely (see Discussion). The results strongly support the argument that the ZAD is not only restricted to the arthropod lineage but is specific for the insect lineage and is likely to have emerged during insect evolution.

3.2.3 Chromosomal distribution and sequence-dependent grouping of ZADs

The positions of ZAD-coding genes in the genome of *D. melanogaster* indicate that they are not randomly distributed on the chromosomes (Figure 3.3). The X chromosome, both arms of the second and the left arm of the third chromosome each contain between 10 and 15 ZAD-coding genes only, whereas the right arm of the third chromosome contains almost half (46 of 98) family members. ZADs can be divided into two large subsets (subsets 1 and 2)

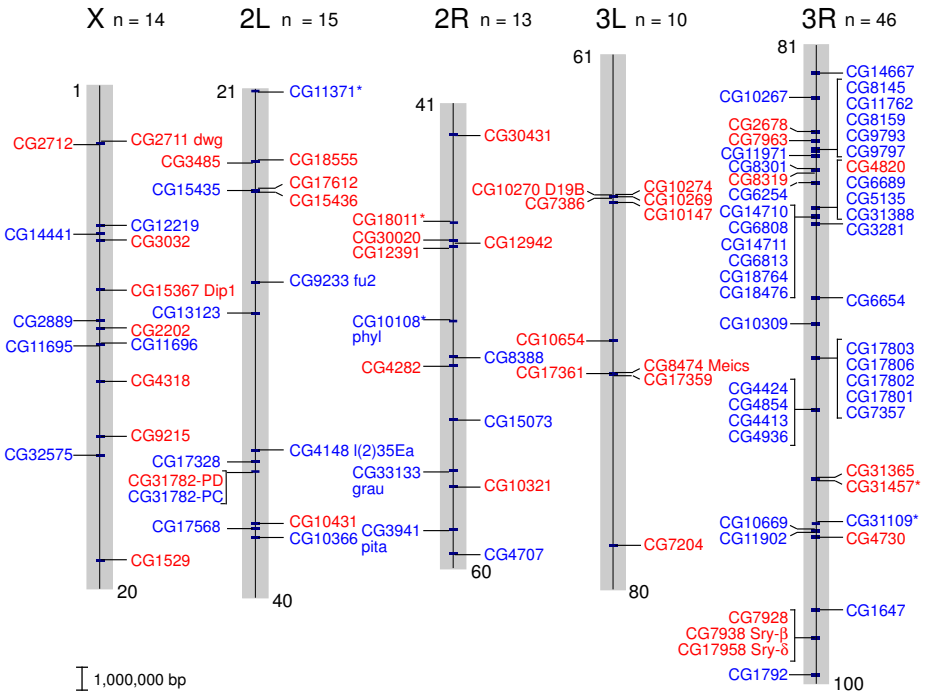


Figure 3.3: Chromosomal distribution of ZAD-coding genes in *D. melanogaster*. In red, single exon ZADs, in blue, two exon ZADs; bp, base pairs; the numbers next to the schematic chromosome or -arm indicate the cytological chromosome bands.

which can be grouped based on the number of exons encoding the ZAD. A total of 41 ZAD coding sequences are encoded by a single exon (subset 1), whereas the open-reading frames of 57 ZADs are split by an intron (subset 2). In most cases except one (see below), the intron is located in a conserved position in block 3 (see Figure B.1).

In order to elucidate the relationships between the ZADs, I constructed a neighbour joining (NJ) tree (Figure 3.4) starting from a multiple sequence alignment (Figure B.1) of all *D. melanogaster* ZADs. The NJ tree was used to determine sequence-related subgroups. The assignment of subgroups is based on the topology of the tree, i.e. sequences were grouped if they can be found on statistical significantly separated branches of the NJ tree. Statistical significance was determined by two tests: (i) a bootstrap test (Sitnikova et al., 1995) that calculates the proportion of times (bootstrap proportion, BP) sequences are grouped when some amino acids in each sequence are randomly substituted in the multiple sequence alignment and new trees are constructed from these "bootstrapped" alignments; (ii) an interior branch test (Sitnikova et al., 1995) that computes the proportion of bootstrap sam-

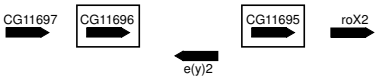
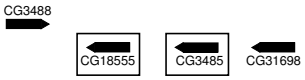
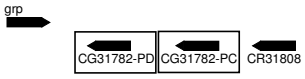
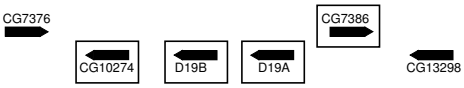
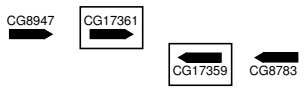
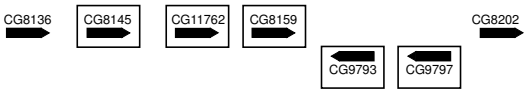
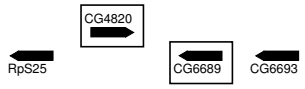
Table 3.1: Sequence related subgroups of *D. melanogaster* ZADs. BP, bootstrap proportion (see main text); CP, confidence probability value of positive branch length (see main text); X, denotes the X chromosome, 2L, 2R, 3L, 3R, denote the left and right chromosome-arms of the second and third chromsosome; N indicates the total number of subgroup members.

Subset	Subgroup	BP	CP	X	2L	2R	3L	3R	**N**
1	A	52	<95	-	6	-	1	-	**7**
1	B	87	99	-	-	-	4	-	**4**
1	C	53	98	2	-	-	-	-	**2**
2	a	76	97	4	-	-	-	-	**4**
2	b	99	99	-	3	-	-	1	**4**
2	c	71	96	-	-	2	-	-	**2**
2	d	<50	96	-	1	-	-	17	**18**
2	e	65	<95	-	-	-	-	3	**3**
2	f	61	98	-	-	-	-	3	**3**
								$\sum$	**47**

ples (see above) producing positive estimates of the branch length (bootstrap confidence probability value, CP value). Sequences were assigned to constitute a subgroup if their BP value was greater than 50 and/or their CP value was greater than 95.

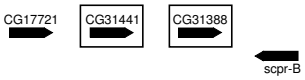
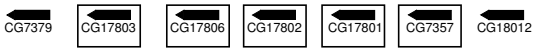
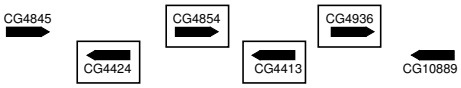
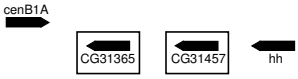
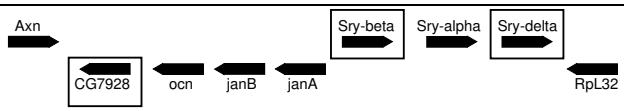
The subgroups contain in most cases either members of subset 1 or 2 exclusively, with the exception of subgroup A (see below). They have been designated as subgroups A-C (subset 1) and a-f (subset 2) and contain 2-18 members (see Table 3.1). The total number of ZAD sequences in subgroups is 47 (Table 3.1).

Table 3.2: Clustered *D. melanogaster* ZAD genes. The arrows indicate the location of the gene on the plus (pointing to the right) and minus (pointing to the left) strand; the clustered ZADs are marked by boxes.

Chromosome -arm	N	χ^2	Gene Map
X	2	1394	
X	3	2151	
2L	2	3449	
2L	2	1642	
2L	2	2212	
2R	2	1036	
3L	4	3398	
3L	2	8089	
3R	5	2218	
3R	2	1403	

continued on next page ...

... continued from previous page

Chromosome -arm	N	χ^2	Gene Map
3R	2	1194	
3R	5	2211	
3R	5	1741	
3R	4	1674†	
3R	2	1255	
3R	2	574	
3R	3	675	

† these four genes do not maximise χ^2 (see Materials and Methods for details)

The sequence similarities between the subgrouped ZADs suggest that they have been generated by gene duplications. This interpretation is consistent with the finding that many ZADs can be found in gene clusters, i.e. a set of genes that are found in close proximity at a genomic locus. Gene clusters have been assigned by a maximum-likelihood method (for details see Materials and Methods). Using this method I found that 49 of the 98 ZAD-coding genes (Table 3.2) are located in gene clusters. 31 out of these 49 clustered genes are also members of subgroups. Of these 31 genes, 24 are found in gene clusters whose members belong exclusively to one subgroup. Other clustered genes can be found as direct neighbours in the NJ tree, e.g. the pair CG17361 and CG17359. However statistical tests performed on this branch did not fulfil the criteria stated above, i.e. BP value <50 and CP value <95. Their grouping is therefore only poorly supported by the statistical tests. They have been consequently excluded from the further analysis. Sequence similarities between the ZADs and their chromosomal clustering suggest that

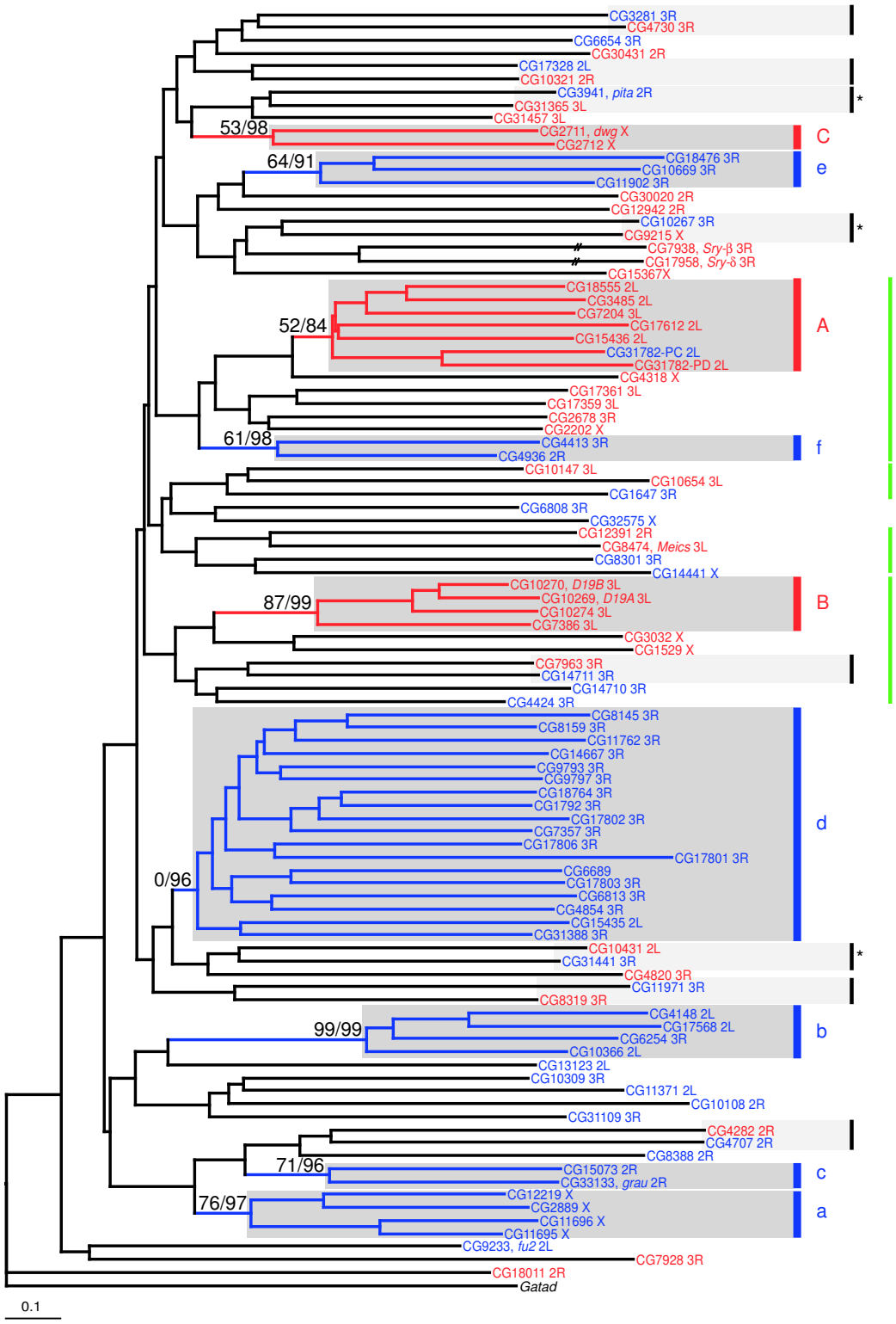


Figure 3.4 *legend on next page*

Figure 3.4: NJ tree of all *D. melanogaster* ZADs. Dark grey boxes indicate subgroups, the subgroup is indicated by the red (single exon ZADs) and blue (two exon ZADs) character next to the red or blue vertical line; the numbers BP/CP denote the bootstrap value and the confidence probability value of the interior branch test (see main text for details) for the proposed subgroups of ZADs, BP/CP values for the other branches have been obmitted to preserve a better representation; light grey boxes indicate neighbours in the tree that belong to subset 1 and 2, respectively, the asteriks marks pairs on different chromosomes; the green vertical lines indicate the groups of ZADs which may have either arisen from multiple retroposition events involving one parental gene, or groups of single exon ZADs that have been generated by local duplication events of a single retroposed gene; in red, single exon ZADs, in blue, two exon ZADs; Gatad denotes a Gata zinc finger used to root the NJ tree; scalebar indicates the number of amino acid substitutions per site.

ZAD-encoding genes have been primarily generated by local gene duplication events. Genes within a sequence-related subgroup that cannot be found in the above mentioned gene clusters can often be found on the same chromosome or even chromosome-arm (Table 3.1), suggesting that they are the result of local gene duplication events followed by para- (not involving the centrosome) or the less frequent occurring pericenctric (involving the centrosome) inversions (reviewed in Powell, 1997), respectively.

Subgroup A is the only exception to the rule that subgroup members belong to either subset 1 or subset 2: one ZAD is encoded by two exons (CG31782-PC) while the ZADs of the six other members of subgroup A are encoded by a single exon (see Table A2). The two ZAD-containing transcripts CG31782-PC and CG31782-PD are not annotated in Release 3.2 genome annotations. It appears likely that they encode for two different genes, but there is no evidence for a gene model for these two genes other than the `genewise` hits to the ZAD HMM. The intron of CG31782-PC is in a unique position and not at the conserved position in block 3 as it can be found in the variable linker region r2 (see Figure 3.1 A). Moreover, a preliminary search in the unassembled whole genome shotgun sequences of four different Drosophila species (*D. simulans*, *D. yakuba*, *D. ananassae* and *D. virilis*) identified orthologous sequences in *D. simulans* and *D. yakuba*. In both putative orthologs the ZAD-coding sequence is not interrupted by an intron (Figure 3.5). Thus, the most parsimonious explanation is that CG31782-PC has gained an intron during evolution.

The NJ tree shows that subsets 1 and 2 are not strictly separated from each other. This observation led to the question whether the genes with two exon ZADs have gained an intron or whether single exon ZADs have lost it. Given the conserved position of the intron in subset 2 ZADs (see above), it seems to be likely that single exon ZADs have lost their introns independently at multiple time points. The loss of an intron has been designated one of the

hallmarks of retroposition events, i.e. the generation of a new (paralogous) gene at a distinct genomic locus by reverse transcription of mRNA from a parental gene (Brosius, 2003). It is, therefore, likely that single exon ZADs have lost their introns due to retroposition events. The observation that the coding sequence of the ZADs of all members of subgroups A-C (except one; see above) are not interrupted by an intron suggests that some of the single exon ZADs have subsequently undergone additional rounds of local gene duplications.

Figure 3.5: Protein and genomic nucleotide sequences of CG31782-PC. Dmel, *D. melanogaster*, Dsim, *D. simulans*, Dyak, *D. yakuba*; in capital letters, protein sequence; yellow boxes, invariant cysteine residues; green capital characters in protein sequence, highly conserved arginine residues; red capital characters in protein sequence, conserved (>60%) hydrophobic amino acid residues; in lowercase letters nucleotide sequence arranged codon-wise top-down; the boundaries of the intron in *D. melanogaster* is indicated in red capital letters.

Retroposition predicts that the newly generated copy (i) is lacking all introns of the parental gene, (ii) has remnants of the polyA-tail of the mRNA following the open reading frame and (iii) is flanked by short duplicated sequences. The NJ tree reveals eight pairs of ZADs, of which one ZAD is encoded by two exons and the other by a single exon. Of these eight pairs, the genes of three pairs can be found on different chromosomes (marked with an asterisk in Figure 3.4), supporting the proposal that they represent retroposed genes. There are four additional cases, which indicate retroposition events. In these cases, ZADs of subset 1 are located on one branch and ZADs of subset 2 on the other that are joined at an internal node (Figure 3.4 marked by green vertical lines). The topology of these branches can be explained by either multiple retroposition events involving the same parental gene or local gene duplications of one ancestral retroposed gene. Based on my analysis, I would predict at least 12 distinct retroposition events. The assignment as retroposed gene, however, has been solely based on the ZAD-coding sequence. 26 of the 41 ZADs of subset 1 have more than one exon coding for the complete amino acid coding-sequence, i.e. although the ZAD

is encoded by one exon the coding sequence outside the ZAD is interrupted by introns. These 12 distinct retroposition events do not have sufficient statistical support from either the BP value or the CP value. Furthermore, the observation that 26 of 41 coding sequences of ZADs of subset 1 contain intronic sequences (see above), indicates that they are not likely to be retroposed genes. However, the lack of statistical support and the presence of introns may be explained by the long time that passed after the proposed retroposition events. Almost all genes involved in the proposed retroposition events, except one (CG4730), can be found in *D. pseudoobscura* (see Table A.2). This finding suggests that these genes have been generated before the lineages of *D. melanogaster* and *D. pseudoobscura* diverged, i.e. the retroposition events must have taken place more than 30 million years ago (see Table 2.2; Figure 3.6). In this light, it is not surprising that hallmarks of retroposition were lost except for the missing intron dividing the ZAD-coding sequence (see also Discussion).

In summary, many ZAD-coding genes in the *D. melanogaster* genome have been most likely generated by local gene duplication events. The distribution of single exon ZADs and two exon ZADs in the NJ tree is consistent with the argument that groups and singletons of ZADs encoded by a single exon emerged independently at distinct time points during evolution involving most likely several independent retroposition events.

3.2.4 Comparison of the ZAD proteome of *D. melanogaster*, *D. pseudoobscura* and *A. gambiae*

The following analysis aims for a comparison of the ZAD proteomes of two other dipteran species, *D. pseudoobscura* and *A. gambiae*, with the ZADs of *D. melanogaster*. These two species have been chosen because their genomes have been sequenced (http://www.hgsc.bcm.tmc.edu/projects/drosophila/; Holt et al., 2002). The availability of whole-genome sequences allowed the *in silico* identification of almost all ZAD-coding genes of these species, which is required to examine both the extent of species-specific expansions and the degree of conservation of ZAD-coding genes that are encoded by the genomes of the three species. Figure 3.6 shows the phylogenetic relationship of these species. *D. melanogaster* and *D. pseudoobscura* diverged approximately 30 million years ago (Table 2.2; Figure 3.6), *D. melanogaster* and *A. gambiae* diverged 250 million years ago (Zdobnov et al., 2002; Figure 3.6). This setting allows a comparison between the ZAD proteomes of two closely related species to the ZAD proteome of a distantly related species.

I was able to identify a total of 115 putative ZAD-coding genes in the genome of *D. pseudoobscura* (Table A.3) and 134 ZAD-coding genes in *A. gambiae* (Table A.4). In *D. pseudoobscura* 103 of the 115 proteins contain

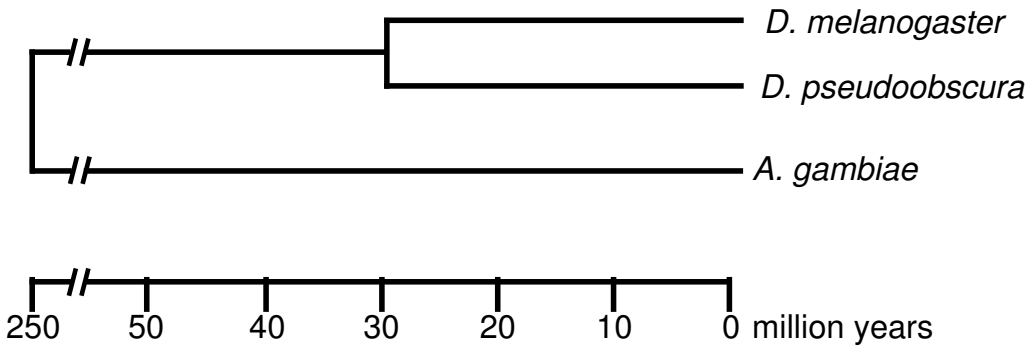


Figure 3.6: Linearised tree indicating the apprimated divergence times of *D. melanogaster*, *D. pseudoobscura* and *A. gambiae*.

a ZAD with associated ZnF domains. In *A. gambiae* 120 of the 134 ZAD-containing proteins also possess ZnF domains. The lack of ZnF motifs in these proteins might be due to three reasons: (i) miss-annotation of these genes resulting in erroneous assignment of the open-reading frame, (ii) the identified sequences correspond to pseudogenes, and (iii) they do not have ZnF domains. The last point was motivated by the finding that in *D. melanogaster* four ZAD-containing proteins have been identified that do not possess ZnF domains (see above). In accordance with this finding, four of the 12 *D. pseudoobscura* ZnF domain lacking proteins are orthologous to the four ZAD-containing genes that lack ZnF motifs in *D. melanogaster*. Even in *A. gambiae*, I was able to identify one orthologous sequence, corresponding to the non-ZFP ZAD-coding gene CG31109 (Table A.2). The lack of ZnF motifs in the remaining proteins might be due to any of the three reasons stated above, since I was not able to distinguish between them in the absence of additional data.

A comparison between the two Drosophilid ZADs revealed that 85 of the *D. melanogaster* ZADs do have at least one ortholog in *D. pseudoobscura* (see Table A.2). That means that some *D. melanogaster* genes have multiple orthologs which have been most likely generated by gene duplication events during the evolution of the lineage that led to *D. pseudoobscura* after the split between the *D. melanogaster* and *D. pseudoobscura* lineages (see below). The evolutionary conservation of the vast majority of *D. melanogaster* ZADs implies that these are functional genes. This has been anticipated already by the finding that for most (81 of 98) *D. melanogaster* ZAD-coding genes at least one EST sequence could be found (see Table A.2). This finding implies that those sequences are transcribed and might act as protein-coding genes.

The 13 and 30 remaining ZAD-coding genes of *D. melanogaster* and *D.*

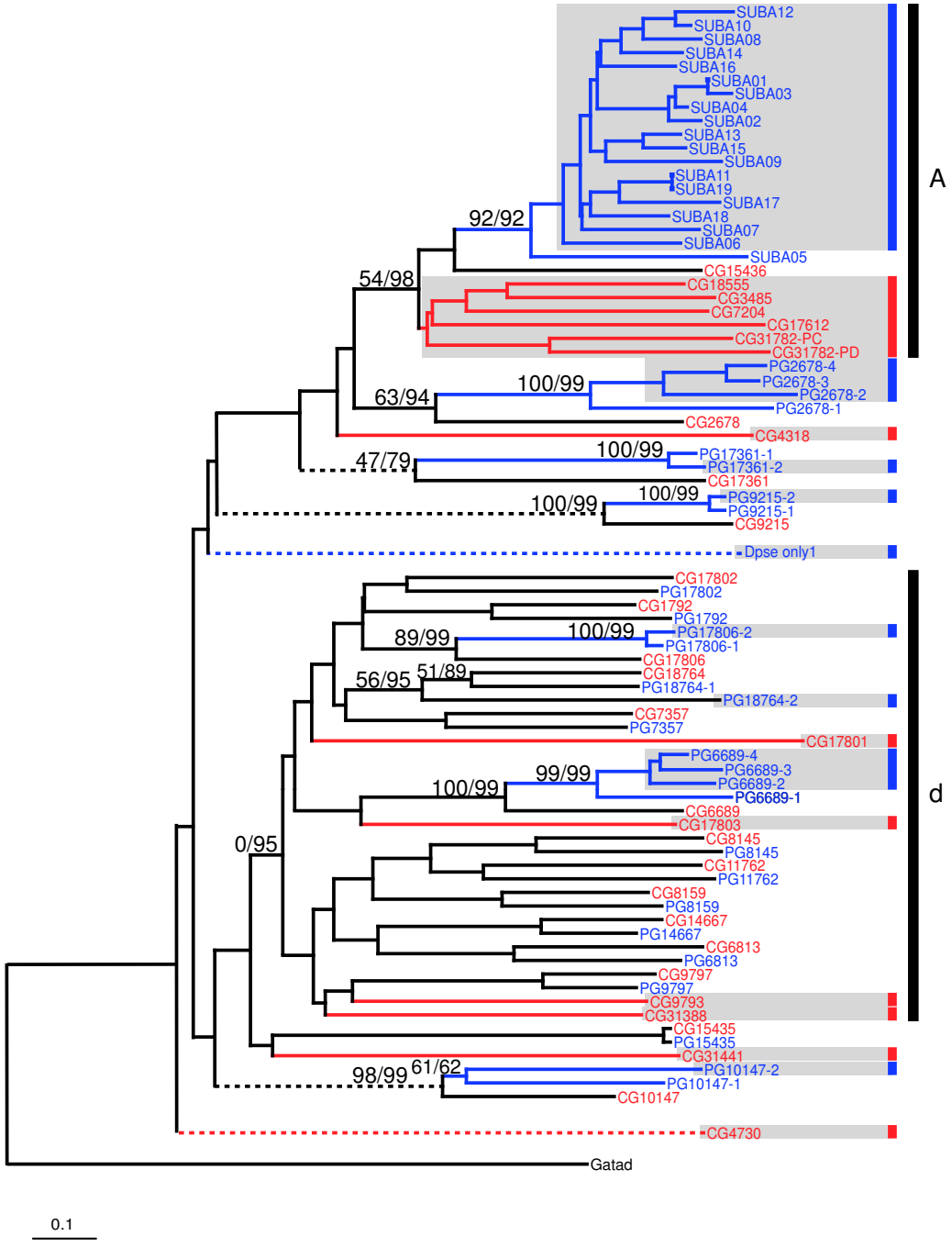


Figure 3.7: NJ tree showing the differentially expanded ZADs in *D. melanogaster* and *D. pseudoobscura*. In red, *D. melanogaster* genes and in blue, *D. pseudoobscura*; grey boxes mark the genes specific for either species; the numbers BP/CP denote the bootstrap value and the confidence probability value of the interior branch test (see main text for details) for the analysed groups; Gatad denotes a Gata zinc finger used to root the NJ tree; the black vertical lines indicate the subgroups A and d, respectively; scalebar indicates the number of amino acid substitutions per site.

pseudoobscura, respectively, represent species-specific genes (Figure 3.7 for an overview). Surprisingly, they are not randomly distributed but mainly confined to two subgroups of *D. melanogaster*, subgroup A and subgroup d (Figure 3.7). Most members of subgroup d can still be found on neighbouring positions in the NJ tree (Figure 3.7). Two genes, however, are missing: CG15435 and CG4854. CG15435 is located on an outgroup branch to subgroup d (Figure 3.7), which demonstrates the close relationship between CG15435 with the members of subgroup d. CG4854, however, is located at a completely different position in the NJ tree (see Figure B.4, marked with an asterisk), suggesting that its grouping to subgroup d in the *D. melanogaster* NJ tree (see Figure 3.4) might have been wrong. This rearrangement is the only example which occurred within subgroups, and seems to be correlated with the low statistical support of subgroup d, i.e. although the CP value (>95) indicates a statistical significant grouping the BP value (<50) does not support the group (see also below). Almost all subgroup d members have orthologs in *D. pseudoobscura*, but for four *D. melanogaster* genes, no orthologous sequences could be identified. In addition, the corresponding genomic loci are massively rearranged, although mapping the neighbours of the missing genes revealed no evidence for sequence gaps, suggesting that the corresponding genes are indeed not present in the *D. pseudoobscura* genome. On the other hand, I was able to identify 5 new subgroup d genes in *D. pseudoobscura*. These have been generated by duplications of the ortholog of CG6689 (four paralogous genes in *D. pseudoobscura*), CG18764 (two paralogs) and CG17806 (two paralogs; Figure 3.7; Table A.2). These results suggest that although most members of the subgroup d have orthologs in *D. pseudoobscura*, there was species-specific expansion of distinct genes within this subgroup to some extent. In subgroup A, a single ortholog pair has been identified, CG15436 and one *D. pseudoobscura* gene (SUBA08 in Figure 3.7), which had one of the neighbouring genes in common with CG15436. In contrast to subgroup d, the remaining 6 members of the *D. melanogaster* subgroup A do not have orthologs in *D. pseudoobscura*. Conversely, 18 subgroup A ZAD-coding genes in *D. pseudoobscura* have no corresponding genes in *D. melanogaster*. The topology of the NJ tree shown in Figure 3.7 strongly suggests that all but the single identified ortholog pair in subgroup A have been newly generated after the *D. melanogaster* and *D. pseudoobscura* lineages have diverged. Both, BP value and CP value support this split between the two species-specific sub subgroups and provide therefore evidence for a *bona fide* example of species-specific expansion involving ZADs.

The evolutionary history of the subgroup A ZADs of *D. melanogaster* and *D. pseudoobscura* could be dated with the help of unassembled whole genome shotgun sequences from four other Drosophila species, *D. simulans*,

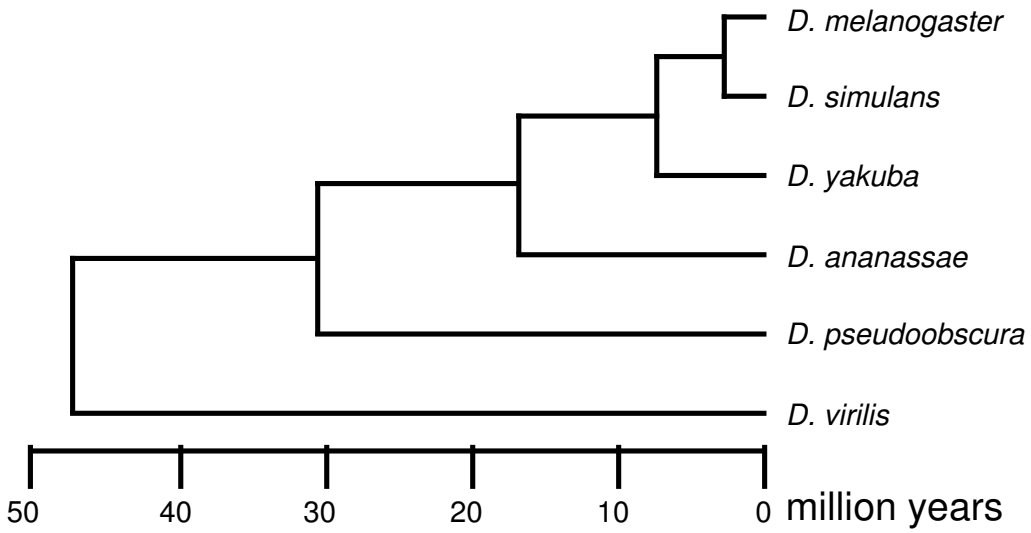


Figure 3.8: Linearised Tree of the analysed Drosophila species. Divergence times was estimated as described in Materials and Methods.

D. yakuba, *D. ananassae* and *D. virilis* (see Figure 3.8 for an overview of the phylogenetic relationships between these species).

A NJ tree constructed with these sequences shows that CG18555, CG3485, CG17612, CG31782-PC and CG31782-PD have orthologous sequences in *D. simulans* and *D. yakuba* (Figure 3.9), indicating that these genes have been generated before the split between *D. yakuba* and the lineage leading to *D. simulans* and *D. melanogaster*. *Neu2* could only be detected in *D. simulans*, indicating that is has been generated before the *D. simulans*/*D. melanogaster* split and after *D. yakuba* diverged (Figure 3.9). Alternatively, it could be that *D. yakuba* has lost *Neu2* after the divergence or I have failed to identify it in the unassembled whole genome shotgun sequences.

In addition, there is one species-specific ZAD sequence that belongs to subgroup A in *D. simulans* and three species-specific subgroup A ZADs in *D. yakuba* (Figure 3.9). *D. ananassae* has no clear orthologous sequences to any subgroup A ZAD of *D. melanogaster* and *D. pseudoobscura* (compare the long branches connecting the *D. ananassae* ZAD sequences with the rest; Figure 3.9). It seems that the four *D. ananassae* ZADs might represent *D. ananassae*-specific ZADs or might be very diverged orthologs of the genes to which they have been assigned to in the NJ tree. Currently, I am not able to distinguish between these two possibilities.

In *D. virilis*, the most diverged species in this comparison, not a single subgroup A gene could be identified (Figure 3.9). This observation implies that the founder gene of this subgroup, CG15436 (see above), has been generated before *D. pseudoobscura* and the *D. melanogaster* have diverged, and most probably after the split between *D. virilis* and the other species oc-

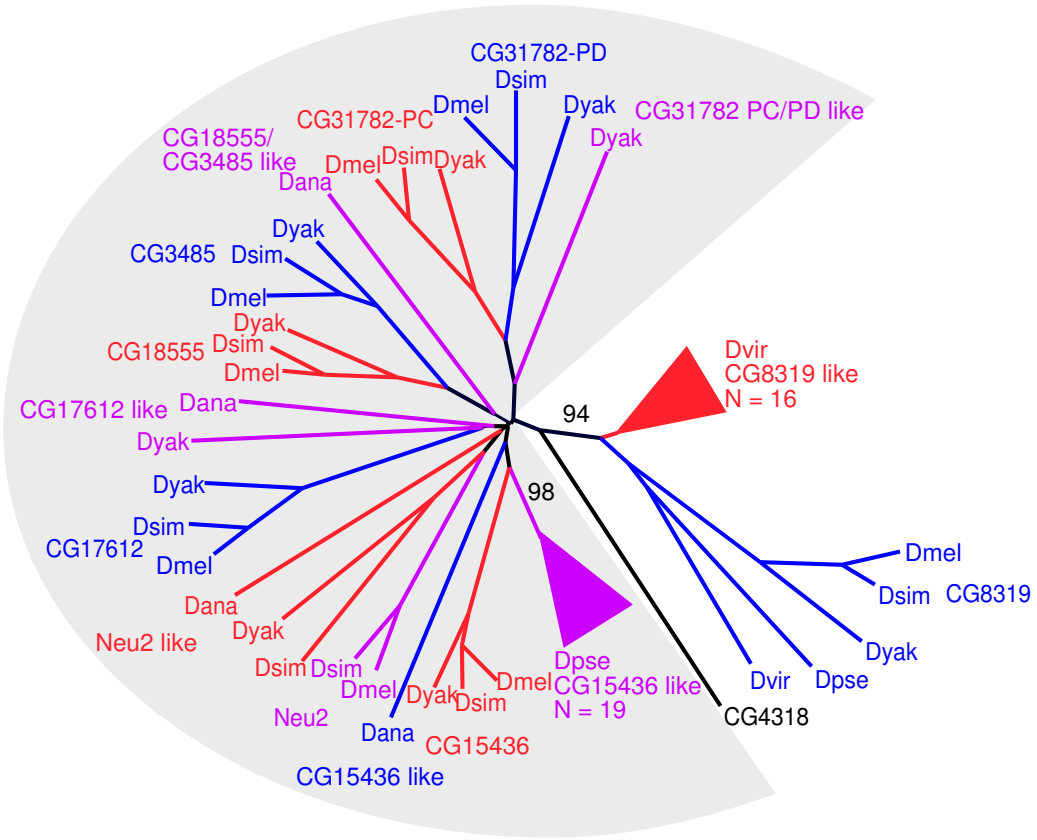


Figure 3.9: Unrooted NJ tree of ZADs of the subgroup A of different Drosophilids. Dmel, *D. melanogaster*; Dsim, *D. simulans*, Dyak; *D. yakuba*, Dpse, *D. pseudoobscura*, Dvir, *D. virilis*; subgroup A is indicated by a gray shadow.

curred. The most probable process that generated CG15436 is a retroposition event (see above), since the coding sequence of CG15436 does not contain an intron. The phylogenetic relationship between the species suggests further that this event must have happened between 30.4 and 47.1 million years ago.

Interestingly, the comparison of the predicted complete protein sequences of the *D. pseudoobscura* subgroup A genes revealed that albeit the ZAD is highly conserved between these paralogous genes, the associated ZnF arrays show only a modest level of conservation. The low level of conservation is also reflected in the different numbers of ZnFs in the corresponding proteins (see Figure 3.10 B). This observation can also be extended to the subgroup A genes in *D. melanogaster* (Figure 3.10 A). The occurrence of the so-called HC-links in the ZnF arrays of all subgroup A members across species lines is indicative for sequence-specific DNA binding (Schuh et al., 1986), suggesting that all members of this subgroup are able to bind DNA in a sequence-specific manner.

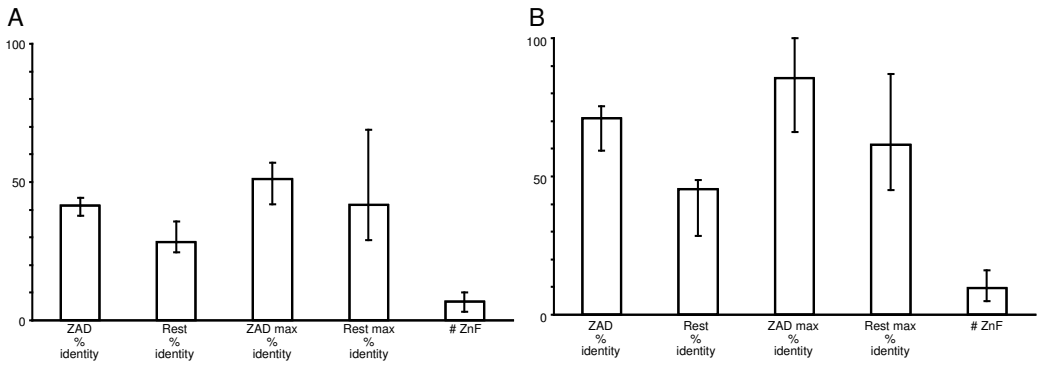


Figure 3.10: Intraspecies comparison of ZADs of subgroup A. (A) *D. melanogaster* ZAD-containing proteins of the subgroup A. The first bar shows the average % identity of the ZAD; the second bar the average % identity of the rest of the protein; the third bar shows the average % identity of the most similar pairs of ZADs for each ZAD-coding sequence; the fourth bar shows the average % identity of the most similar pairs of the remainder of the proteins; the last bar shows the average number of ZnF motifs per protein; the errorbars indicate the minimal and maximal observed value. (B) *D. pseudoobscura* ZAD-containing proteins of the subgroup A. See A for details.

A possible explanation for the differential level of similarity is that the selective pressure that constrains the sequence for the ZAD is higher than the need to preserve the ZnF arrays. Given that the ZAD is predicted to be a protein-protein interaction domain for homodimer formation and/or heterodimer association of closely related ZADs, it is likely that the higher constraint for the ZAD is due to the preservation of the ability to form heterodimers between the family members. In this view, conservation of the ZAD together with divergence of the ZnF arrays would combine the ability to form heterodimers with different DNA specificities. This feature in turn, might be key to combinatorial regulation of gene expression by the ZAD-containing ZFPs, provided that they are transcriptional regulators.

Comparison of the genomes of the Drosophilids *D. melanogaster* and *D. pseudoobscura* revealed that the majority of ZAD-coding genes are conserved between these two species. This finding implies that the majority of them carry important functions. The differences between these two ZAD proteomes are mainly due to differential expansion of the subgroup A members (a total of 6 of 13 genes specific for *D. melanogaster* and 18 of 30 genes specific for *D. pseudoobscura*). This finding shows that a divergence time of approximately 30 million years seems to be sufficient to acquire significant differences in the repertoire of proteins that are potentially involved in the regulation of transcription.

The differences become more apparent in a comparison of the ZAD proteome of *D. melanogaster* and *A. gambiae*. Here, I was able to identify only six orthologous pairs which had sufficient support from either BP value or CP value (see Figure 3.11). The orthology could be further verified by the addition of *D. pseudoobscura* ZAD sequences to the protein alignment, which did not change the statistical support significantly. Apart from these six conserved genes, there are large groups of proteins which are species-specific (see Figure 3.11). Almost all subgroups defined by the NJ tree involving only *D. melanogaster* ZADs (Figure 3.4) showed no rearrangements. The only exception is subgroup d, like in the *D. melanogaster* and *D. pseudoobscura* analysis (see above). Subgroup d seems to be sensitive to the addition of sequences to the NJ tree, which may be correlated to the high number of genes in this subgroup. Nevertheless, the core, i.e. 16 of 18 sequences remain clustered, justifying the assignment as a subgroup of sequence-related ZADs. The clear separation of most *A. gambiae* (93 of 134) and *D. melanogaster* (75 of 98) ZADs on distinct branches suggests that the majority of these genes have been generated after the divergence of the *A. gambiae* and *D. melanogaster* lineages. This finding, together with the low number of orthologs (see also Discussion), confirm that the ZAD-containing genes are indeed subject to species- and lineage-specific expansions.

3.2.5 Hints towards an involvement of ZAD-coding genes in speciation

Small interaction domains that mediate protein-protein association or protein-nucleic acid-binding, such as the ZAD and the ZnF domain, are prevalent in proteins which are subject to lineage-specific or even species-specific expansion. This prevalence might be due to the relative stability of the fold of these small domains (Lespinet et al., 2002) and by the versatile functional spectrum such domains add to a protein. Interactions mediated by these domains might establish a mutual dependency of two or more genomic loci for their proper functions. Examples for such mutual dependencies are the formation of higher order protein complexes which are functional only when all binding partners are present or the interaction of transcription factors with their target cis-regulatory elements, a process necessary for the transcriptional regulation of nearby genes. These interactions might be subject to divergent co-adaptive evolution in isolated subgroups of populations, which in turn leads to subgroup-specific alleles of pairs, triplets etc. of genes that are incompatible with the alleles of the other nascent species. Matings between individuals of nascent species lead to offspring whose allelic combination is not compatible anymore. This effect is called hybrid incompatibility. Such incompatibilities are thought to be one of the first steps in the forma-

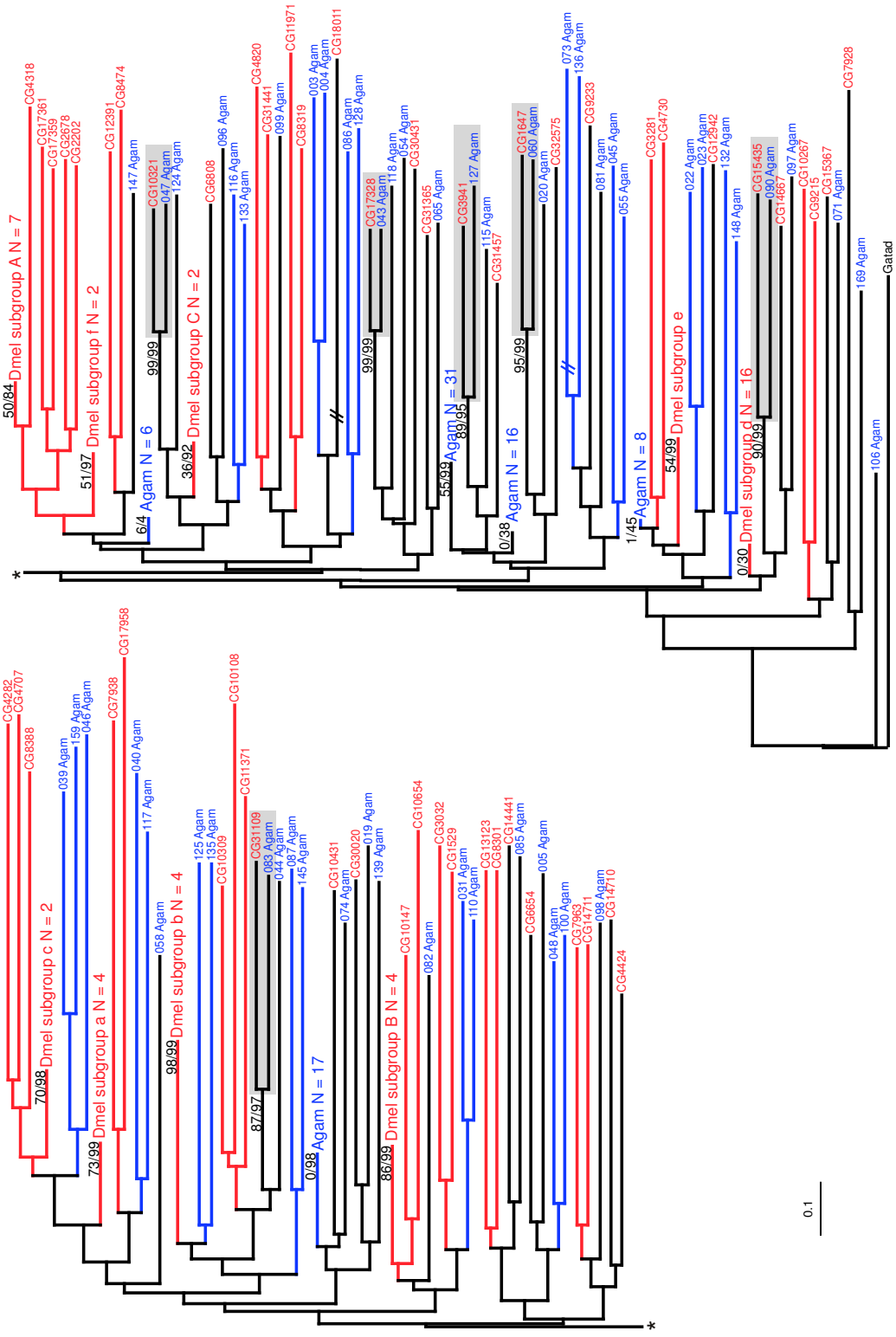


Figure 3.11 *legend on next page*

Figure 3.11: NJ tree showing *D. melanogaster* and *A. gambiae* ZADs. In red, *D. melanogaster* ZADs, in blue *A. gambiae* ZADs; grey boxes indicate orthologous pairs; the numbers BP/CP denote the bootstrap value and the confidence probability value of the interior branch test (see main text for details) for orthologous pairs and the species-specific branches; Gatad denotes a Gata zinc finger used to root the NJ tree; scalebar indicates the number of amino acid substitutions per site.

tion of reproductive barriers that reinforce and contribute to the formation of new species. This model of speciation has been referred to as DM model (Dobzhansky, 1936; Muller, 1939; Muller, 1940). A consequence of this model is that some genomic loci, the so-called "speciation genes", which build up inter-species incompatibilities, will not be exchanged between the nascent species any longer, while the vast majority of genes might still be able to cross the species boundary upon secondary contact. This model predicts that "speciation genes" have diverged significantly earlier than the bulk of the genome (Wu, 2001; Wu and Ting, 2004).

The ZAD is a protein-protein interaction domain that can mediate the formation of homo- and heterodimers. It is therefore possible that some sets of ZADs divergently co-evolve in the sense described above and lead to incompatibilities in hybrid offspring. Given that the ZAD-coding genes frequently duplicate within the genome, it is possible that relaxed evolutionary constraints after the duplication event result in a rapid divergence of one copy. This copy might then acquire new interaction partners and might force co-adaption to fix the interaction. Another explanation might be that the ZnF arrays change their DNA specificity, this in turn might lead to co-evolution of cis-regulatory elements, which has been noted earlier (Bonneton et al., 1997; Ruez et al., 1998).

To test the hypothesis that ZAD-coding genes are "speciation genes", I estimated the divergence time of orthologous pairs of protein-coding genes on a genome-wide scale between *D. melanogaster* and *D. pseudoobscura*. The estimation of the divergence time has been based on the rate of synonymous exchanges (dS). Synonymous exchanges, which do not lead to changes in the amino acid sequence of the protein, are thought to be neutral and exhibit the behaviour of a molecular clock, i.e. the calculated rate of synonymous exchanges correlates with the divergence time. This correlation is due to the assumption that the real dS is approximately constant, i.e. larger dS values indicate a longer divergence time than smaller dS values.

For the genome-wide comparison of protein-coding genes between *D. melanogaster* and *D. pseudoobscura* the dS was calculated with the help of the PAML program (version 3.2; Yang, 1997; for details see Materials and Methods). In this approach, I calculated the dS for 11,099 orthologous pairs of *D. melanogaster* and *D. pseudoobscura*. The data derived by these calculations

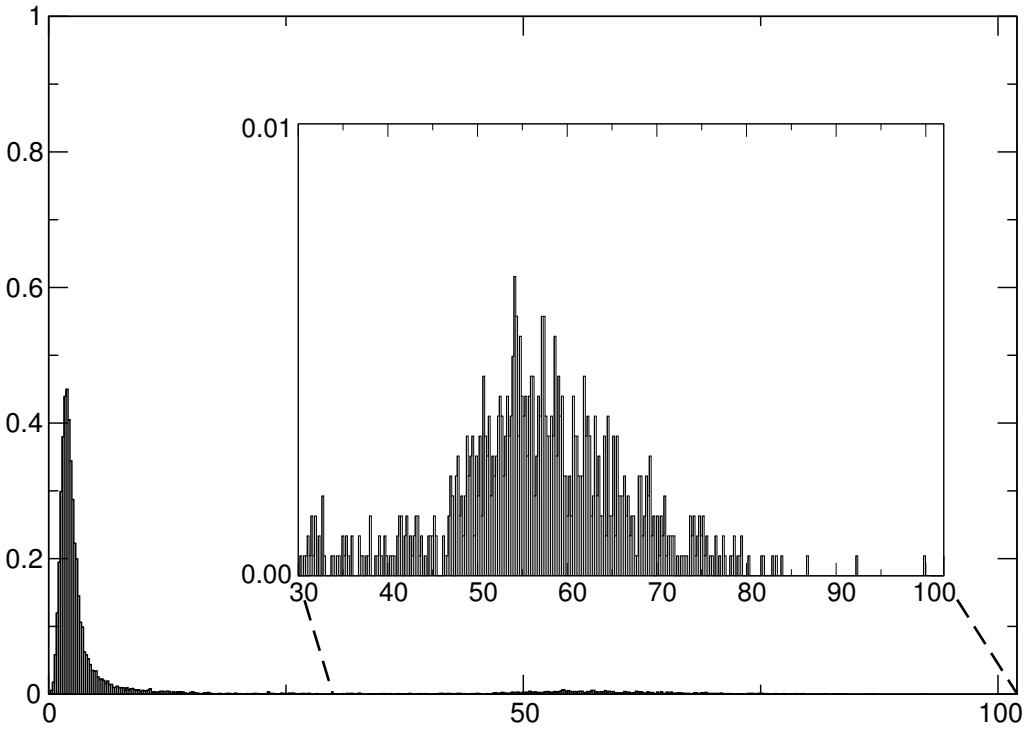


Figure 3.12: Histogramm representation of the distribution of dS values calculated after Goldman and Yang (1994). On the x-axis dS values in synonymous exchanges per synonmous site and on the y-axis the frequency. The dS value range > 30 has been blown up to better visualise the second distribution (see insert).

have been visualised by a histogram representation (Figure 3.12). Significantly, the distribution is not monodispersed but shows a second "peak" around dS=50, indicating that there are indeed two classes of genes: the vast majority 10,298 (93%) can be found at values $dS < 40$, only 801 (7%) had dS values greater than 40 (Figure 3.12 insert).

Interestingly, among the 801 genes with $dS > 40$, 14 ZAD-coding genes can be found, corresponding to 16.5% of the 85 orthologous ZAD-coding gene pairs. This number is significantly higher than expected (χ^2=10.08; df=1; p < 0.002). $dS > 40$ indicates that these 14 ZAD-coding genes have diverged earlier than the bulk of genes within the genome. These findings are the basis of the hypothesis that at least some ZAD-coding genes are "speciation genes" and may contribute to the formation of reproductive barriers and, in turn, to speciation (see Discussion).

Chapter 4

Discussion

4.1 C2H2 zinc finger proteins in the *D. melanogaster* genome

I identified a total of 454 ZFPs encoded by 359 genes, which corresponds to about 2.6% of the protein coding genes of the genome. The number of 359 ZFP-coding genes is very similar to the 357 ZFP-coding genes identified by Lander et al. (2001) in the *D. melanogaster* genome. The same study showed that the genes coding for the ZFP-family constitute in fact the most populous protein family in the *D. melanogaster* proteome. Together, these results indicate that the ZFP-family clearly represents the most abundant nucleic acid-binding protein family in *D. melanogaster*, followed by the homeobox-containing protein family with 148 genes (=1.1% extrapolated to the whole genome; Lander et al., 2001). Only 31.7% (see 3.1) of these genes seem to be evolutionary conserved at the sequence level in vertebrates and/or nematodes, whereas 68.3% (see 3.1) seem to be specific for the arthropod crown group represented by *D. melanogaster*. Similar findings have been reported for all transcriptional regulator families (Coulson and Ouzounis, 2003). This study showed that almost half of all transcriptional regulator protein families are not shared between the major eukaryotic lineages (Coulson and Ouzounis, 2003). The low level of conservation may be explained by the participation of ZFP-coding genes in lineage-specific expansions (Lespinet et al., 2002).

The in-depth analysis of *D. melanogaster*-specific ZFPs led to the identification of the ZAD, a domain of 71-97 amino acids. The domain is almost exclusively associated with ZnF motifs. The 94 ZFPs that contain this newly identified domain constitute the single largest subfamily of *D. melanogaster* ZFPs. This subfamily has undergone lineage-specific expansions and accounts for 94 out 245 *D. melanogaster*-specific ZFP-coding genes, which corresponds to 38.7%. The BTB/POZ domain-containing subfamily of ZFPs is the other large group of *D. melanogaster*-specific ZFPs. This subfamily has 15 members, i.e. 6.1% of all arthropod-specific ZFP-coding genes. These two groups together account for nearly half, i.e 44.8%, of all *D. melanogaster*-specific ZFP-coding genes. On the protein level, there are 95 ZAD-ZFPs and 50 BTB/POZ-ZFPs, corresponding to 30.6% and 16.1% of the *D.*

melanogaster-specific ZFPs, respectively. This means that the BTB/POZ-ZFP-coding genes are in contrast to the ZAD-ZFP-coding genes subject to a high level of alternative splicing and/or alternative promotor usage. The N-terminal BTB/POZ domain coding sequences are fused to varying C-terminal coding-sequences including ZnF arrays.

4.2 The zinc finger associated domain

The association between the ZAD and ZnF domains is not exclusive. There are at least four proteins in the *D. melanogaster* genome that contain a ZAD but no ZnFs or other DNA binding domains. A similar finding has been reported for the SCAN domain which is specific for vertebrates (Collins et al., 2001). The SCAN domain has been proposed to mediate homo- and/or heterodimer formation (Schumacher et al., 2000) and is often associated with ZnF motifs (Collins et al., 2001). In analogy to the ZAD, some SCAN domain-containing proteins have been identified that do not contain ZnF domains (Castillo et al., 1999; Sander et al., 2000; Sander and Morris, 2002). For example, the PPAR γ coavtivator-2 (PGC-2) encodes a SCAN domain protein and lacks other diagnostic protein domains. It has been shown to interact with the orphan nuclear receptor Peroxisome Proliferator-Activated Receptor γ (PPAR γ). The association between PGC-2 and PPAR γ increases the transcriptional activation activity of PPAR γ (Castillo et al., 1999). One striking example for a potential function of a ZAD without ZnF domains is the protein encoded by *phyl* (Chang et al., 1995; Dickson et al., 1995). Phyl contains a ZAD-like motif and has been shown to interact with a transcriptional repressor encoded by *tramtrack* and a ring finger protein encoded by *seven in absentia*. The interaction between the three proteins has been suggested to lead to the targeted degradation of Tramtrack by the ubiquitin-dependent protein degradation pathway (Li et al., 1997; Tang et al., 1997). These examples suggest that isolated SCAN domains and ZADs represent protein-protein interaction motifs which participate in modulating the biological activity of transcription factors at the posttranscriptional level.

4.2.1 The ZAD is insect-specific

The initial finding that the ZAD is refined to a group of lineage-specific ZFPs suggested that they may be specific for the arthropod lineage. However, the subsequent analysis of the phylogenetic distribution of the ZAD showed that it is further restricted to insects in the arthropod lineage. The results where obtained by searches in EST databases, which are not as biased towards known protein motifs as protein databases. The lack of ZAD-coding sequence in the vertebrate and nematode portion of the EST database is highly significant. The absence of the domain in EST sequences of the sister

groups of the insects, *Crustacea* and *Chelicerata*, where only >20,000 and >10,000 EST sequences, respectively, were screened, is less convincing. It is still possible that sequences encoding ZAD-like domains are present, in very low numbers, in the genomes of other arthropods. Nevertheless, the results clearly indicate that if not the ZAD per se, at least the expansion of the ZAD-family of ZFPs is restricted to the insect order. A definite answer whether the domain can be found in the sister groups of the insects can ultimately be answered once whole-genome sequence data of species of the crustacean and cheliceratan orders are available.

4.2.2 Properties of the ZAD

The structure of the ZAD of Grauzone revealed, as predicted, an independently folding zinc-binding protein module, which mediates the formation of homodimers and/or heterodimers of closely related ZADs (Jauch et al., 2003). This function is consistent with the finding that 44 ZAD-containing proteins physically interact with each other in yeast cells, including seven proteins that form homodimers, as has been revealed by large scale yeast two hybrid screens (Giot et al., 2003). In contrast to the initial prediction that ZADs can associate with closely related family members only, Giot et al. (2003) showed that at least in yeast cells ZAD-containing proteins can interact with a wide variety of other ZAD-containing proteins, i.e. an average of 3.1 interactions per protein, and a maximum number of 15 interactions (CG7386). One has to note, however, that although the ZAD was shown to mediate such interactions, it cannot be excluded that ZAD-containing proteins interact via their ZnF domain or via other unidentified protein motifs present in those interacting proteins.

The ZAD and the expansion of ZAD-coding genes can be considered as lineage-specific, at least for the arthropod lineage. This is reminiscent to the SCAN and KRAB domains, which are restricted to the vertebrate lineage and also participate in lineage-specific expansions (Collins et al., 2001; Lander et al., 2001; Lespinet et al., 2002). The ZAD as well as the SCAN domain most probably mediate interactions between other ZADs and SCAN domains, respectively. The KRAB domain has been proposed to be a transcriptional repressor domain (Urrutia, 2003). As the two vertebrate-specific protein-protein interaction domains, the ZAD is also associated with ZnF motifs and can be found in a large number of lineage-specific ZFPs. The high number of ZAD-, SCAN- and KRAB-encoding genes in the genomes of *D. melanogaster* and humans, respectively, suggests that the combination of protein-protein interaction modules with ZnF motifs is beneficial for the individual organism and may play a substantial role in the combinatorial control of gene activities.

4.2.3 Evolutionary aspects of the ZAD

The detailed analysis of the ZAD-family of genes in *D. melanogaster* suggests that new ZAD-coding genes are mainly generated by local duplication events which were sometimes followed by para- or pericentric inversions (reviewed in Powell, 1997). In addition, I have identified two distinct subgroups of ZAD-coding genes containing one or no intron in the ZAD-coding sequence, respectively. This notion led to the question whether the genes with two exon ZADs have gained an intron or whether single exon ZADs have lost it. Interestingly, in all but one of the two exon ZADs, the intron is precisely at the same position within the coding sequence of the ZAD (see Figure B.1). Since it is unlikely that a gain of intron sequences always occurs in the same position, as evidenced by the single example where the intron is found in a different position, it appears straightforward to assume that the single exon ZADs must have lost their intron.

A generally accepted mechanism for the loss of intron sequences is a retroposition event in which the spliced and processed mRNA copy of the primary transcript is retrotranscribed and integerates into a new chromsomal site (reviewed in Brosius, 2003). A model that explains the duplication/retroposition event that results in the generation of a new ZAD gene at a different locus is shown in Figure 4.1. This model explains the lack of separation of ZADs of subset 1 and subset 2 in the NJ tree, as it allows for an independent repeated generation of intron-less copies of parental intron-containing genes. The observation that 26 of 41 ZAD-coding genes contain introns in their coding sequences outside the ZAD-coding sequence, argues against the proposed retroposition event. The lack of direct statistical support for these events and the occurrence of introns outside the ZAD-coding sequence can be attributed to the long period the retroposed genes were present in the genome, as has been revealed by the identification of orthologs for almost all of these genes in the *D. pseudoobscura* genome. This suggests that the events that led to the generation of these genes date back at least 30 million years, the approximated divergence time of the two lineages. The relative old age of these retroposed copies explains the lack of additional hallmarks of retroposition other than the loss of the intron in the ZAD-coding sequence, as has been noted for other retroposition events (Betrán et al., 2002). Betrán et al. (2002) have shown that only the youngest gene (<15 million years) in their collection retained the direct repeats flanking the insertion and that the polyA tail rapidly degenerates.

Retroposition might have played a role in the loss of the intron in the ZAD-coding sequence. If true, the retroposition events might have been followed by the gain of introns outside the ZAD-coding sequence in 26 cases.

Alternatively, the introns might have been generated by the recruitment of nearby exons of other genes (Long et al., 2003). Retroposition seems to be the most parsimonious explanation for the loss of the intron in the ZAD-coding sequences for several reasons: (i) the two exon ZAD-coding sequences most likely correspond to the ancestral state (see above); (ii) the loss of intronic sequences is not accompanied with a significant amount of insertions or deletions surrounding the ancestral intron position (Figure B.1), which would be expected for losses of intronic sequences in the genomic DNA; (iii) in three of the eight pairs of ZADs, where one ZAD is encoded by a single exon and the other by two exons, that are likely to represent retroposition events the coding sequence for the whole protein is contained in a single exon. In sumary, the most likely mechanism that led to the loss of the intron in ZADs of subset 1 is a retroposition event, since other, alternative explanations, like the deletion of intronic sequences in genomic DNA seem to be very unlikely.

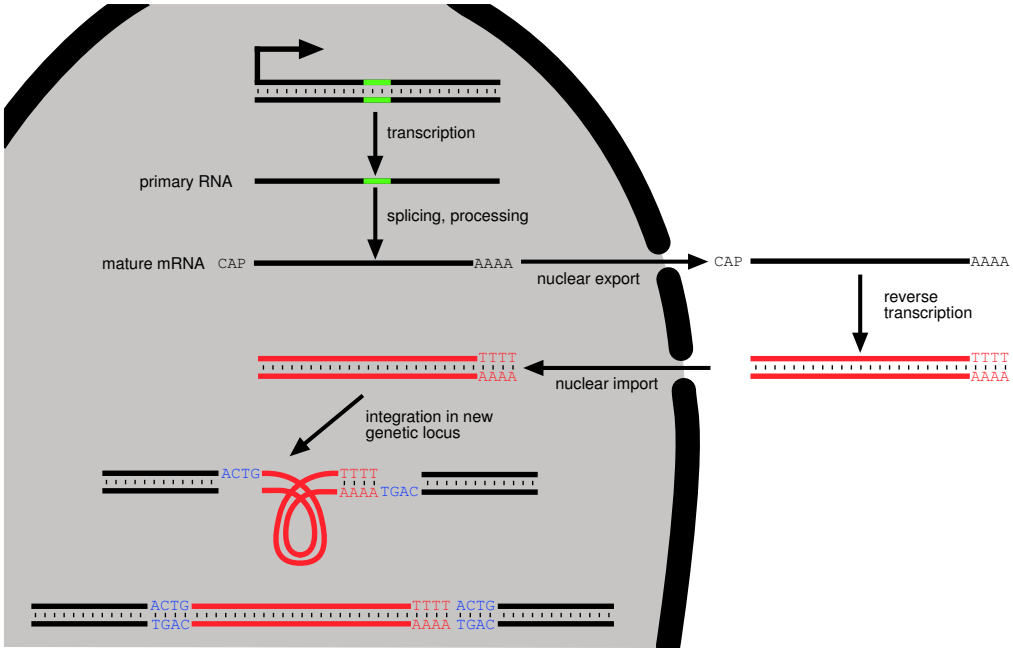


Figure 4.1: Retroposition mechanism. Intronic sequences are marked in green; `AAAA` indicates the poly-adenine tail; `CAP` represents the 3'-cap structure.

I was able to place 47 ZADs in nine different subgroups. Members of these subgroups are highly similar at the protein sequence level of the ZAD. For some members of subgroups the sequence similarities extend towards the ZnF arrays, i.e. they might also have very similar DNA binding affinities.

These genes might therefore have redundant functions. Redundant or at least partially redundant functions of ZADs may explain why the majority of ZAD-coding genes have escaped functional detection by mutagenesis screens (e.g. Nüsslein-Volhard and Wieschaus, 1980; Peter et al., 2002; Spradling et al., 1999).

Another explanation for the lack of identified mutant alleles might be that a mutation in these genes do not cause easily scoreable phenotypes but lead to rather subtle effects on the fitness of individuals carrying mutant alleles of these genes. The identification of 85 orthologous gene pairs in *D. melanogaster* and *D. pseudoobscura* suggests that the majority of the ZAD-coding genes indeed carry functions which are essential. Thus, they have been conserved after the divergence of the two lineages leading to *D. melanogaster* and *D. pseudoobscura*, for a period of 30 million years. On the other hand, a comparison between *D. melanogaster* and *A. gambiae* genomes revealed that only six orthologous pairs could be assigned, i.e. a total of only 6.1% of the *D. melanogaster* ZAD-encoding genes. This number is much lower than recent results obtained by a whole-genome comparison between *D. melanogaster* and *A. gambiae*. These studies revealed that 44% of the *D. melanogaster* genes have orthologous sequences in *A. gambiae* (Zdobnov et al., 2002). The roughly eight-fold difference in the percentages of all versus ZAD-coding genes can in part be attributed to the more stringent criteria used to identify orthologous pairs in the subset of ZAD-coding genes. A comparison of datasets generated by the same methods, i.e. a reciprocal BLAST search (Zdobnov et al., 2002; see Materials and Methods) between the ZAD-proteome of the two species, revealed 10 orthologous genes pairs, i.e. 10.2%. This marginal increase in the number of conserved genes does not significantly change the result that the percentage of conserved ZAD-coding genes is significantly lower than the percentage of conserved genes in a whole-genome comparison (χ^2=24.9, df=1, p< 0.0005).

It is therefore not surprising to find that the majority of ZAD-coding genes of both species can be found on branches that are specific for the respective species. This finding suggests that the majority of genes coding for ZAD-containing proteins have been generated after the divergence of the lineages leading to *A. gambiae* and *D. melanogaster*, respectively. This interpretation of the branching pattern of the NJ tree is consistent with the proposal that ZAD-coding genes participate in lineage-specific expansions. It cannot be excluded that the evolutionary history of the two compared ZAD proteomes involved a rapid loss of ZAD-coding genes encoded by the genome of the last common ancestor of *D. melanogaster* and *A. gambiae*. Gene loss might have accompanied the evolution of individual subfamilies of ZAD-coding genes, but seems to be unlikely to have contributed significantly to the evolution

of the whole ZAD family of genes. Such a scenario would require that the genome of the last common ancestor of *D. melanogaster* and *A. gambiae* coded for a significant higher number of ZAD-coding genes.

Four of the five ZAD-coding genes of *D. melanogaster*, which have been identified by mutant alleles, are not conserved in *A. gambiae*. Mutations in these four genes lead to a lethal phenotype. This observation, suggests that at least some lineage-specific ZAD-coding genes may have acquired vital functions and become indispensable for the organism. These four genes and in fact most of the other *D. melanogaster* ZAD-coding genes have orthologous counterparts in *D. pseudoobscura*. This observation suggests that they arose after the divergence of *A. gambiae* and the Drosophilid lineages and before the split leading to *D. pseudoobscura* and *D. melanogaster*. The proposed orthology of these genes, suggests that they are likely to have the same function in the two species.

A survey of transposon insertions into ZAD-coding genes of *D. melanogaster* (Bellen et al., 2004; Peter et al., 2002; Thibault et al., 2004) showed that most of the targeted genes do not have a lethal phenotype (20 of 23 targeted genes). These 23 genes together with the five genes mentioned above sum up to 29 loci of which eight are mutatable to lethality, i.e. 27.6% of the assayed loci. This percentage is very similar to an approximation of the percentage of lethal loci of 30% (Miklos and Rubin, 1996), indicating that the ZAD-coding genes do not behave differently from the rest of the genome.

It seems that some ZAD-coding genes have acquired essential functions during evolution, as has been revealed by the lack of orthologous sequences in *A. gambiae* for four of the five experimentally characterised genes. The sequence conservation of these genes in *D. pseudoobscura* suggests that they might have acquired their functions before *D. melanogaster* and *D. pseudoobscura* diverged. There are little differences in the ZAD-proteome of *D. melanogaster* and *D. pseudoobscura*, the major difference is generated by the differential expansion of ZADs of subgroup A. There is some evidence for differential expansion of this subgroup in at least three different Drosophila species. While the differential expansions in *D. melanogaster* and *D. pseudoobscura* can be unambiguously shown, the relationship between the ZADs of subgroup A of *D. ananassae* and those of *D. melanogaster* and *D. pseudoobscura* is unclear. The intra-species comparison of the ZAD-coding genes of subgroup A in *D. melanogaster* and *D. pseudoobscura* revelealed that the ZAD is much better conserved than the remainder of the protein. This might be due to an increased selective constraint on the ZAD-coding sequence to preserve the ability to form heterodimers between closely related ZADs. The ability to form heterodimers combined with different DNA-specificities provided by the varying ZnF arrays might be key to combinatorial regulation

of gene expression. In this view the generation of many different BTB/POZ domain-containing ZFPs from a single gene by alternative splicing and/or alternative promotor usage (see above) might be viewed as an alternative way to combine different DNA-specificities with the ability to form higher order complexes.

Strikingly, there is no positive evidence for subgroup A members in the *D. virilis* genome. Keeping in mind that unassembled whole genome shotgun sequences are not the best choice to infer the absence of genes in a particular genome, it seems very unlikely that the failure to find any ZAD of subgroup A in the *D. virilis* genome is due to gaps in the coverage of genomic loci, since I was able to find many other ZAD-coding sequences. Transposon insertions in 3 of the 7 *D. melanogaster* ZADs of subgroup A are reported to be homozygous viable. This suggests that the proteins encoded by these genes are not essential for viability. This conclusion is consistent with the finding that they are not present in *D. virilis*, since *D. virilis* can cope with the lack of these genes very well. Given the conservation of the ZADs of subgroup A of *D. melanogaster* over a period of 7.3 million years, as evidenced by the identification of orthologous sequences in *D. simulans* and *D. yakuba*, it seems unlikely that the presence of these genes is neutral with respect to evolution. The estimated half-life of a duplicated gene in *D. melanogaster* is approximately 3.2 million years (Lynch and Conery, 2003). These genes are older than it is expected for neutral gene duplications, which is consistent with the proposal that the genes of subgroup A possess functions that, although not essential for viability, are conserved in evolution.

4.2.4 Hints towards an involvement of ZAD-coding genes in speciation

Lineage-specific restriction and expansion of protein families have been described for many functional classes of proteins over different phylogenetic distances (for example Fortna et al., 2004; Lespinet et al., 2002; Mark et al., 1999; Shannon et al., 2003). These proteins are often characterised by a certain domain architecture, where a sequence comparison of some domains led to the identification of lineage-specific novelties and expansion (Lespinet et al., 2002). These domains often include sequence motifs that are thought to mediate protein-protein interactions and protein-nucleic acid interactions, like the KRAB, SCAN, ZAD or ZnF domains. The massive increase of the overall number of these domains and in particular the number of domains within individual proteins from the unicellular yeast *Saccharomyces cervisea* to humans has been suggested to reflect the overall increase in organismal complexity (Lander et al., 2001). In contrast, a correlation between the number of genes and the organismal complexity seems to be less strict (Lander

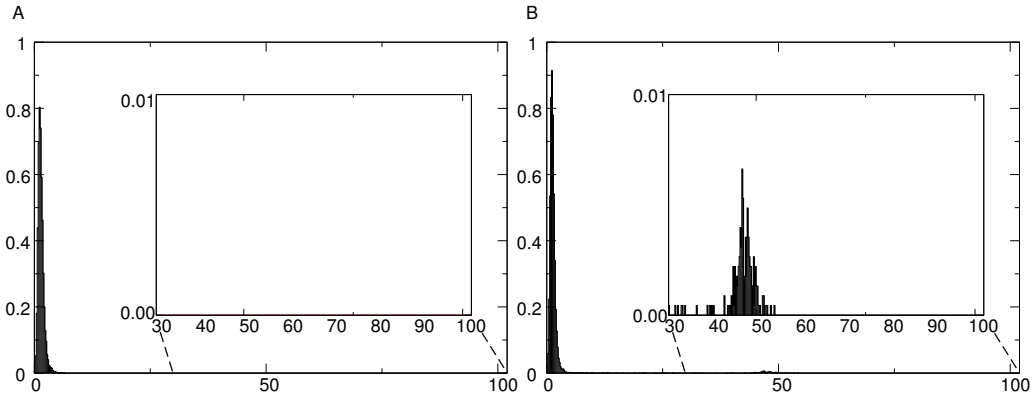


Figure 4.2: Histogramm representation of the distribution of dS values calculated after Nei and Gojobori (1986) and Goldman and Yang (1994) using fixed codon frequencies. On the x-axis dS values and on the y-axis the frequency. **A** distribution of dS calculated after Nei and Gojobori (1986), N = 10414. **B** distribution of dS calculated after Goldman and Yang (1994) using fixed codon frequencies,N = 11,099; there are 173 genes in the second distribution $dS > 40$.

et al., 2001). This suggests that an increase of interaction domains causes an increase in the complexity and possibly multifunctionality of higher order protein-nucleic acid complexes.

This phylogenetic trend, however, cannot explain the differential expansion of gene families in species with small phylogenetic distances, since their organismal complexity is very similar, as revealed in the comparison of the ZAD proteomes of *D. melanogaster* and *D. pseudoobscura*. Differential expansions of protein families of closely related species have also been reported earlier. For example, a comparison between the human chromosome 19 with related sequences in the mouse genome shows independent expansion of genes containing the KRAB domain (Mark et al., 1999; Shannon et al., 2003). The functional implications of these expansions are not clear and have not been addressed experimentally.

An approximation of the divergence time of all putative ortholog pairs of *D. melanogaster* and *D. pseudoobscura* revealed a diverse group of 801 genes that seem to have diverged significantly earlier than the bulk of the genome. In this group, ZAD-coding genes have been found to be statistical significantly overrepresented. These results have been obtained by the use of maximum-likelihood methods implemented in PAML (Goldman and Yang, 1994) to estimate the rate of synonymous exchanges dS. The second distribution around $dS = 50$ (Figure 3.12), vanishes if I use the method proposed by Nei and Gojobori (1986) implemented in PAML (see Figure 4.2 A). The latter method, however, neglects the transition/transversion ratio and the

Table 4.1: Distribution of the different dS class genes on the chromosomes or chromosome arms. X, X chromosome; 2L, left arm of the second chromosome; 2R, right arm of the second chromosome; 3L, left arm of the third chromosome; 3R, right arm of the third chromosome; 4, fourth chromosome; U, unassigned scaffolds.

Class	X	2L	2R	3L	3R	4	U
$dS \leq 40$	1516	1852	2100	2009	2666	65	90
$dS > 40$	232	125	147	136	154	1	6
	=13.3%	=6.3%	=6.5%	=6.3%	=5.4%	=1.5%	6.2%

codon-usage bias, which leads to a severe bias in the estimate of dS, leading to a underestimtation of dS (Yang and Bielawski, 2000). Bierne and Eyre-Walker (2003), on the other hand, showed that the method of Goldman and Yang (1994) which is used by PAML has to be taken with caution, since in some applications the Goldman and Yang (1994) method leads to erroneous results, like the failure to detect a negative correlation between codon usage bias and the rate of synonymous exchanges (Bierne and Eyre-Walker, 2003). Unfortunately, the second distribution around $dS = 50$ seems to be dependent on the estimation of the codon usage bias, since an analysis conducted with an equal probability for all codons (= 1/61) reveals only a very reduced second distribution (Figure 4.2 B).

The percentage of orthologous gene pairs that are in the $dS > 40$ class is roughly two times higher on the X chromosome than on the autosomes (13.3 % on the X chromosome versus 6 % on the autosomes; Table 4.1). Many studies (e.g. Orr and Coyne, 1989) have shown that X chromosomal linked genes have the largest effect on hybrid sterility and inviability. For example, a comparison between the number of loci leading to hybrid male sterility in crosses between *D. simulans* and *D. mauritiana* on the X and the third chromosome revealed that the X chromosome has a 2.5 times higher density of hybrid male sterility factors than the autosomes (Tao et al., 2003). The results of Tao et al. (2003) are therefore in good agreement with the results I obtained in my analyses. Taken together my observations are in favor of the argument that at least some of the genes and possibly also some of the ZAD-coding genes are among those which are referred to as "speciation genes".

Chapter 5

Summary

Whole genome sequence data helps to understand the molecular evolution of proteins and the underlying DNA sequences. I performed an in-depth *in silico* analysis of C2H2 zinc finger proteins (ZFPs) in *Drosophila melanogaster* and identified a subfamily which is characterised by a N-terminal domain of 71-97 amino acids ("zinc finger associated domain"; ZAD). The *Drosophila melanogaster* genome contains 94 genes encoding members of this subfamily and additional four genes that code for proteins containing an isolated ZAD. The members of this subfamily constitute the single largest group of ZFPs, i.e. 26.1% of a total of 359 ZFP coding genes..

The ZAD represents an independently folding domain that mediates homodimer formation and probably association of closely related ZAD family members. Sequence comparison and the X-ray cristallographic structure of the ZAD of the transcription factor Grauzone revealed four invariant cysteine residues that mediate binding to a zinc ion.

The ZAD-coding genes fall into two large subsets: ZADs of subset 1 are encoded by a single exon and ZADs of subset 2 by two exons. Analysis of these two subsets led to the proposal that subset 1 ZADs have lost their intron at multiple time points by separate retroposition events.The majority of ZAD-coding genes were generated by local gene duplication events.

Comparison of the ZAD proteomes of *Drosophila melanogaster* with the closely related Drosophilid *Drosophila pseudoobscura* and the mosquito *Anopheles gambiae* revealed that ZAD-containing proteins are subject to lineage-specific expansion. The degree of expansion correlates with the phylogenetic divergence of the species.

A whole-genome comparison of all orthologous pairs of *Drosophila melanogaster* and *Drosophila pseudoobscura* identified two classes of genes: a small class (7%) likely to have diverged significantly earlier than the other orthologs. Within this fraction of the genome ZAD-coding genes are over-represented. A recent proposal stated that the so-called "speciation genes", which represent the molecular basis of reproductive isolation, have diverged earlier than the vast majority of genes. In this light, ZAD-coding genes might represent "speciation genes".

Chapter 6

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., et al. (2000). The genome sequence of *Drosophila melanogaster*. Science **287**, 2185-2195.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**, 3389-3402.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**, 796-815.

Aravind, L., Watanabe, H., Lipman, D. J., and Koonin, E. V. (2000). Lineage-specific loss and divergence of functionally linked genes in eukaryotes. Proc Natl Acad Sci USA **97**, 11319-11324.

Bardwell, V.J., and Treisman, R. (1994). The POZ domain: a conserved protein-protein interaction motif. Genes Dev **8**, 1664-1677.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., et al. (2004). The Pfam protein families database. Nucleic Acids Res **32**, D138-141.

Bellefroid, E.J., Poncelet, D.A., Lecocq, P.J., Revelant, O., and Martial, J.A. (1991). The evolutionarily conserved Krüppel-associated box domain defines a subfamily of eukaryotic multifingered proteins. Proc Natl Acad Sci USA **88**, 3608-3612.

Bellen, H. J., Levis, R. W., Liao, G., He, Y., Carlson, J. W., Tsang, G., Evans-Holm, M., Hiesinger, P. R., Schulze, K. L., Rubin, G. M., et al. (2004). The BDGP Gene Disruption Project: Single Transposon Insertions Associated With 40% of Drosophila Genes. Genetics **167**, 761-781.

Berg, J. M., and Shi, Y. (1996). The galvanization of biology: a growing appreciation for the roles of zinc. Science **271**, 1081-1085.

Betrán, E., Thornton, K., and Long, M. (2002). Retroposed new genes out of the X in Drosophila. Genome Res **12**, 1854-1859.

Bhaskar, V., Valentine, S. A., and Courey, A. J. (2000). A functional interaction between dorsal and components of the Smt3 conjugation machinery. J Biol Chem **275**, 4033-4040.

Bierne, N., and Eyre-Walker, A. (2003). The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. Genetics **165**, 1587-1597.

Birney, E., Thompson, J. D., and Gibson, T. J. (1996). PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. Nucleic Acids Res **24**, 2730-2739.

Bonneton, F., Shaw, P. J., Fazakerley, C., Shi, M., and Dover, G. A. (1997). Comparison of *bicoid*-dependent regulation of *hunchback* between *Musca domestica* and *Drosophila melanogaster*. Mech Dev **66**, 143-156.

Brosius, J. (2003). The contribution of RNAs and retroposition to evolutionary novelties. Genetica **118**, 99-116.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol **268**, 78-94.

Castillo, G., Brun, R. P., Rosenfield, J. K., Hauser, S., Park, C. W., Troy, A. E., Wright, M. E., and Spiegelman, B. M. (1999). An adipogenic cofactor bound by the differentiation domain of PPARgamma. EMBO J **18**, 3676-3687.

The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282**, 2012-2018.

Celniker, S. E., Wheeler, D. A., Kronmiller, B., Carlson, J. W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S. P., Frise, E., et al. (2002). Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. Genome Biol **3**, RESEARCH0079.

Chang, H. C., Solomon, N. M., Wassarman, D. A., Karim, F. D., Therrien, M., Rubin, G. M., and Wolff, T. (1995). *phyllopod* functions in the fate determination of a subset of photoreceptors in Drosophila. Cell **80**, 463-472.

Chen, B., Harms, E., Chu, T., Henrion, G., and Strickland, S. (2000). Completion of meiosis in Drosophila oocytes requires transcriptional control by Grauzone, a new zinc finger protein. Development **127**, 1243-1251.

Chung, H. R., Schäfer, U., Jäckle, H., and Böhm, S. (2002). Genomic expansion and clustering of ZAD-containing C2H2 zinc-finger genes in Drosophila. EMBO Rep **3**, 1158-1162.

Collins, T., Stone, J. R., and Williams, A. J. (2001). All in the family: the BTB/POZ, KRAB, and SCAN domains. Mol Cell Biol **21**, 3609-3615.

Coulson, R. M., and Ouzounis, C. A. (2003). The phylogenetic diversity of eukaryotic transcription. Nucleic Acids Res **31**, 653-660.

Crozatier, M., Kongsuwan, K., Ferrer, P., Merriam, J. R., Lengyel, J. A., and Vincent, A. (1992). Single amino acid exchanges in separate domains of the Drosophila *serendipity delta* zinc finger protein cause embryonic and sex biased lethality. Genetics **131**, 905-916.

Dickson, B. J., Dominguez, M., van der Straten, A., and Hafen, E. (1995). Control of Drosophila photoreceptor cell fates by *phyllopod*, a novel nuclear protein acting downstream of the Raf kinase. Cell **80**, 453-462.

Dobzhansky, T. (1936). Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. Genetics **21**, 113-135.

Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics **14**, 755-763.

Fahmy, O. G., and Fahmy, M. (1959). New mutants report. Dros Inf Serv **33**, 82-94.

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., et al. (2004). Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol. **2**, E207.

Gaszner, M., Vazquez, J., and Schedl, P. (1999). The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction. Genes Dev **13**, 2098-2107.

Gemünd, C., Ramu, C., Altenberg-Greulich, B., and Gibson, T. J. (2001). Gene2EST: a BLAST2 server for searching expressed sequence tag (EST) databases with eukaryotic gene-sized queries. Nucleic Acids Res **29**, 1272-1277.

Gilbert, S.F. (2003). Developmental Biology. (Sunderland, Sinauer Associated).

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., et al. (2003). A protein interaction map of *Drosophila melanogaster*. Science **302**, 1727-1736.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science **274**, 546, 563-547.

Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11**, 725-736.

Harms, E., Chu, T., Henrion, G., and Strickland, S. (2000). The only function of Grauzone required for Drosophila oocyte meiosis is transcriptional activation of the *cortex* gene. Genetics **155**, 1831-1839.

Hilfiker, A., Hilfiker-Kleiner, D., Pannuti, A., and Lucchesi, J. C. (1997). *mof*, a putative acetyl transferase gene related to the Tip60 and MOZ human genes and to the SAS genes of yeast, is required for dosage compensation in Drosophila. EMBO J **16**, 2054-2060.

Hoch, M., Schröder, C., Seifert, E., and Jäckle, H. (1990). cis-acting control elements for *Krüppel* expression in the Drosophila embryo. EMBO J **9**, 2587-2595.

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M., Wides, R., et al. (2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. Science **298**, 129-149.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD - Visual Molecular Dynamics. J Molec Graphics **14**, 33-38.

Huynh, K. D., and Bardwell, V. J. (1998). The BCL-6 POZ domain and other POZ domains interact with the co-repressors N-CoR and SMRT. Oncogene **17**, 2473-2484.

Jauch, R., Bourenkov, G. P., Chung, H. R., Urlaub, H., Reidt, U., Jäckle, H., and Wahl, M. C. (2003). The zinc finger-associated domain of the Drosophila transcription factor Grauzone is a novel zinc-coordinating protein-protein interaction module. Structure **11**, 1393-1402.

Klug, A., and Schwabe, J. W. (1995). Protein motifs 5. Zinc fingers. FASEB J **9**, 597-604.

Koonin, E. V., Aravind, L., and Kondrashov, A. S. (2000). The impact of comparative genomics on our understanding of evolution. Cell **101**, 573-576.

Kraulis, P.J. (1991). MOLSCRIPT: A Program to Produce Both Detailed and Schematic Plots of Protein Structures. J Appl Cryst **24**, 946-950.

Krishna, S. S., Majumdar, I., and Grishin, N. V. (2003). Structural classification of zinc fingers: survey and summary. Nucleic Acids Res **31**, 532-550.

Kumar, S., Tamura, K., Jakobsen, I. B., and Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17**, 1244-1245.

Kumar, S., Tamura, K., Nei, M. (2004). MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform **5**, 150-163.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature **409**, 860-921.

Laundrie, B., Peterson, J. S., Baum, J. S., Chang, J. C., Fileppo, D., Thompson, S. R., and McCall, K. (2003). Germline cell death is inhibited by P-element insertions disrupting the *dcp-1/pita* nested gene pair in Drosophila. Genetics **165**, 1881-1888.

Lespinet, O., Wolf, Y. I., Koonin, E. V., and Aravind, L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res **12**, 1048-1059.

Li, S., Li, Y., Carthew, R. W., and Lai, Z. C. (1997). Photoreceptor cell differentiation requires regulated proteolysis of the transcriptional repressor Tramtrack. Cell **90**, 469-478.

Long, M., Deutsch, M., Wang, W., Betrán, E., Brunet, F.G., and Zhang, J. (2003). Origin of new genes: evidence from experimental and computational analyses. Genetica **118**, 171-182.

Lynch, M., and Conery, J. S. (2003). The evolutionary demography of duplicate genes. J Struct Funct Genomics **3**, 35-44.

Mark, C., Abrink, M., and Hellman, L. (1999). Comparative analysis of KRAB zinc finger proteins in rodents and man: evidence for several evolutionarily distinct subfamilies of KRAB zinc finger genes. DNA Cell Biol **18**, 381-396.

Matyash, A., Chung, H. R., and Jäckle, H. (2004). Genome-wide mapping of in vivo targets of the Drosophila transcription factor *Krüppel*. J Biol Chem **279**, 30689-30696

McGinnis, W., Garber, R.L., Wirz, J., Kuroiwa, A., and Gehring, W.J. (1984). A homologous protein-coding sequence in Drosophila homeotic genes and its conservation in other metazoans. Cell. **37**, 403-408.

Merika, M., and Thanos, D. (2001). Enhanceosomes. Curr Opin Genet Dev **11**, 205-208.

Merritt, E.A., and Murphy, M.E.P. (1994) Raster3D Version 2.0 - A Program for Photorealistic Molecular Graphics. Acta Cryst **D50**, 869-873.

Miklos, G. L., and Rubin, G. M. (1996). The role of the genome project in determining gene function: insights from model organisms. Cell **86**, 521-529.

Miller, J., McLachlan, A. D., and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. EMBO J **4**, 1609-1614.

Morgenstern B. (1999). DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics **15**, 211-218.

Muller, H. J. (1939). Reversibility in evolution considered from the standpoint of genetics. Biol Rev Camb Philos Soc **14**, 261-280.

Muller, H. J. (1940). Bearing of the Drosophila work on systematics. In The New Systematics, H. Huxley, ed. (Oxford, Clarendon Press), pp. 185-268.

Murre, C., McCaw, P.S., and Baltimore, D. (1989). A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. Cell **56**, 777-783.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol **3**, 418-426.

Nüsslein-Volhard, C., and Wieschaus, E. (1980). Mutations affecting segment number and polarity in Drosophila. Nature **287**, 795-801.

Orr, H. A., and Coyne, J. A. (1989). The genetics of postzygotic isolation in the *Drosophila virilis* group. Genetics **121**, 527-537.

Page, R. D. M. (1996). TREEVIEW: An application to display phylogenetic trees on personal computers. CABIOS **12**, 357-358

Pankratz, M. J., Busch, M., Hoch, M., Seifert, E., and Jäckle, H. (1992). Spatial control of the gap gene *knirps* in the Drosophila embryo by posterior morphogen system. Science **255**, 986-989.

Pavletich, N. P., and Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science **252**, 809-817.

Payre, F., Buono, P., Vanzo, N., and Vincent, A. (1997). Two types of zinc fingers are required for dimerization of the serendipity delta transcriptional activator. Mol Cell Biol **17** , 3137-3145.

Payre, F., Crozatier, M., and Vincent, A. (1994). Direct control of transcription of the Drosophila morphogen Bicoid by the Serendipity delta zinc finger protein, as revealed by in vivo analysis of a finger swap. Genes Dev **8**, 2718-2728.

Payre, F., Noselli, S., Lefrere, V., and Vincent, A. (1990). The closely related Drosophila Sry beta and Sry delta zinc finger proteins show differential embryonic expression and distinct patterns of binding sites on polytene chromosomes. Development **110**, 141-149.

Peter, A., Schöttler, P., Werner, M., Beinert, N., Dowe, G., Burkert, P., Mourkioti, F., Dentzer, L., He, Y., Deak, P., et al. (2002). Mapping and identification of essential gene functions on the X chromosome of Drosophila. EMBO Rep **3**, 34-38.

Pi, H., Huang, S. K., Tang, C. Y., Sun, Y. H., and Chien, C. T. (2004). *phyllopod* is a target gene of proneural proteins in Drosophila external sensory organ development. Proc Natl Acad Sci USA **101**, 8378-8383.

Pi, H., Wu, H. J., and Chien, C. T. (2001). A dual function of *phyllopod* in Drosophila external sensory organ development: cell fate specification of sensory organ precursor and its progeny. Development **128**, 2699-2710.

Powell, J. R. (1997). Progress and Prospects in Evolutionary Biology: The Drosophila Model (Oxford, Oxford University Press).

Rosenberg, U. B., Schröder, C., Preiss, A., Kienlin, A., Côté, S., Riede, I., and Jäckle, H. (1986). Molecular genetics of *Krüppel*, a gene required for segmentation of the Drosophila embryo. Nature **313**, 336-339.

Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol **266**, 525-539.

Ruez, C., Payre, F., and Vincent, A. (1998). Transcriptional control of Drosophila *bicoid* by *Serendipity delta*: cooperative binding sites, promoter context, and co-evolution. Mech Dev **78**, 125-134.

Russo, C.A., Takezaki, N., and Nei, M. (1995). Molecular phylogeny and divergence times of drosophilid species. Mol Biol Evol **12**, 391-404.

Rutherford, K., Parkhill, J. Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., and Barrell, B. (2000) Artemis: sequence visualisation and annotation. Bioinformatics **16**, 944-945.

Sander, T. L., Haas, A. L., Peterson, M. J., and Morris, J. F. (2000). Identification of a novel SCAN box-related protein that interacts with MZF1B. The leucine-rich SCAN box mediates hetero- and homoprotein associations. J Biol Chem **275**, 12857-12867.

Sander, T. L., and Morris, J. F. (2002). Characterization of the SCAN box encoding RAZ1 gene: analysis of cDNA transcripts, expression, and cellular localization. Gene **296**, 53-64.

Sayle, R.A., and Milner-White, E.J. (1995). RASMOL: biomolecular graphics for all. Trends Biochem Sci **20**, 374.

Schuh, R., Aicher, W., Gaul, U., Côté, S., Preiss, A., Maier, D., Seifert, E., Nauber, U., Schröder, C., Kemler, R., and Jäckle, H. (1986). A conserved family of nuclear proteins containing structural elements of the finger protein encoded by *Krüppel*, a Drosophila segmentation gene. Cell **47**, 1025-1032.

Schumacher, C., Wang, H., Honer, C., Ding, W., Koehn, J., Lawrence, Q., Coulis, C. M., Wang, L. L., Ballinger, D., Bowen, B. R., and Wagner, S. (2000). The SCAN domain mediates selective oligomerization. J Biol Chem **275**, 17173-17179.

Schüpbach, T., and Wieschaus, E. (1989). Female sterile mutations on the second chromosome of *Drosophila melanogaster*. I. Maternal effect mutations. Genetics **121**, 101-117.

Scott, M.P., and Weiner, A.J. (1984). Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila. Proc Natl Acad Sci USA **81**, 4115-4119.

Shannon, M., Hamilton, A. T., Gordon, L., Branscomb, E., and Stubbs, L. (2003). Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. Genome Res **13**, 1097-1110.

Shiu, S. H., and Li, W. H. (2004). Origins, lineage-specific expansions, and multiple losses of tyrosine kinases in eukaryotes. Mol Biol Evol **21**, 828-840.

Sitnikova, T., Rzhetsky, A., and Nei, M. (1995). Interior-branch and bootstrap tests of phylogenetic trees. Mol Biol Evol **12**, 319-333.

Spradling, A. C., Stern, D., Beaton, A., Rhem, E. J., Laverty, T., Mozden, N., Misra, S., and Rubin, G. M. (1999). The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. Genetics **153**, 135-177.

Stathopoulos, A., Van Drenth, M., Erives, A., Markstein, M., and Levine, M. (2002). Whole-genome analysis of dorsal-ventral patterning in the Drosophila embryo. Cell **111**, 687-701.

Tang, A. H., Neufeld, T. P., Kwan, E., and Rubin, G. M. (1997). PHYL acts to down-regulate TTK88, a transcriptional repressor of neuronal cell fates, by a SINA-dependent mechanism. Cell **90**, 459-467.

Tao, Y., Chen, S., Hartl, D. L., and Laurie, C. C. (2003). Genetic dissection of hybrid incompatibilities between *Drosophila simulans* and *D. mauritiana*. I. Differential accumulation of hybrid male sterility effects on the X and autosomes. Genetics **164**, 1383-1397.

Thibault, S. T., Singer, M. A., Miyazaki, W. Y., Milash, B., Dompe, N. A., Singh, C. M., Buchholz, R., Demsky, M., Fawcett, R., Francis-Lang, H. L., et al. (2004). A complementary transposon tool kit for *Drosophila melanogaster* using P and piggyBac. Nat Genet **36**, 283-287.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22**, 4673-4680.

Urrutia, R. (2003). KRAB-containing zinc-finger repressor proteins. Genome Biol **4**, 231.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. Science **291**, 1304-1351.

Williams, A.J., Blacklow, S.C., and Collins, T. (1999) The zinc finger-associated SCAN box is a conserved oligomerization domain. Mol Cell Biol **19**, 8526-8535.

Wolfe, S. A., Nekludova, L., and Pabo, C. O. (2000). DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct **29**, 183-212.

Wu, C. I. (2001). The genic view of the process of speciation. J Evol Biol **14**, 851-865.

Wu, C. I., and Ting, C. T. (2004). Genes and speciation. Nat Rev Genet **5**, 114-122.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13**, 555-556.

Yang, Z., and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. Trends Ecol Evol **15**, 496-503.

Zdobnov, E. M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R. R., Christophides, G. K., Thomasova, D., Holt, R. A., Subramanian, G. M., et al. (2002). Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. Science **298**, 149-159.

Zollman, S., Godt, D., Prive, G.G., Couderc, J.L., and Laski F.A. (1994). The BTB domain, found primarily in zinc finger proteins, defines an evolutionarily conserved family that includes several developmentally regulated genes in Drosophila. Proc Natl Acad Sci USA **91**, 10717-10721.

Appendix A

Tables

Table A.1: All *Drosophila melanogaster* ZFPs. # ZnF, number of ZnF motifs found; C, *Caenorhabditis elegans*; M, *Mus musculus*.

Symbol	Protein	# ZnF	conserved in		additional domains
			C	M	
		Chromosome X			
CG17829	CG17829-PA	10	x	x	-
CG17829	CG17829-PC	9	x	x	-
CG11398	CG11398-PA	7	-	-	-
br	CG11491-PA	2	-	-	BTB
br	CG11491-PB	2	-	-	BTB
br	CG11491-PD	2	-	-	BTB
br	CG11491-PE	2	-	-	BTB
br	CG11491-PG	2	-	-	BTB
CG14050	CG14050-PA	1	x	-	-
dwg	CG2711-PA	8	-	-	ZAD
CG2712	CG2712-PA	8	-	-	ZAD
CG3526	CG3526-PB	1	-	-	ZZ
CG3526	CG3526-PA	1	-	-	ZZ
CG33221	CG33221-PA	5	-	-	-
CG6121	CG6121-PA	1	x	x	MOZ/SAS
peb	CG12212-PA	13	-	x	-
Bteb2	CG2932-PA	3	-	x	-
CG12682	CG12682-PA	1	-	-	-
CG32772	CG32772-PA	8	-	-	AT_hook
ovo	CG6824-PB	4	x	x	-
ovo	CG6824-PC	4	x	x	-
ovo	CG6824-PA	4	x	x	-
CG32767	CG32767-PA	7	-	-	-
mof	CG3025-PA	1	x	x	MOZ/SAS
CG12236	CG12236-PA	2	-	-	BTB
CG12236	CG12236-PB	2	-	-	BTB
CG12219	CG12219-PA	4	-	-	ZAD
CG3847	CG3847-PA	4	-	-	-
CG3224	CG3224-PA	1	x	x	-
CG14441	CG14441-PA	2	-	-	HTH_7, ZAD
CG14438	CG14438-PA	18	-	-	GoLoco
CG14435	CG14435-PA	1	x	x	zf-C3HC4
CG3032	CG3032-PA	9	-	-	ZAD

Table A.1 *continued on next page*

Table A.1 *continued from previous page*

Symbol	Protein	# ZnF	conserved in		additional
			C	M	domains
CG9650	CG9650-PA	5	x	x	-
CG9650	CG9650-PB	5	x	x	-
CG9650	CG9650-PC	6	x	x	-
CG32719	CG32719-PA	1	-	-	-
CG2116	CG2116-PA	7	-	-	-
CG2120	CG2120-PA	7	-	-	-
CG2129	CG2129-PA	10	-	-	-
CG15336	CG15336-PA	4	-	-	-
CG10959	CG10959-PA	9	-	-	-
CG18262	CG18262-PA	8	-	-	-
CG12772	CG12772-PA	1	-	x	-
CG7065	CG7065-PA	1	-	-	-
Dip1	CG15367-PA	1	-	-	ZAD
btd	CG12653-PA	3	-	-	-
Sp1	CG1343-PA	3	-	x	-
Sp1	CG1343-PB	3	-	x	-
CG2889	CG2889-PA	10	-	-	ZAD
CG9817	CG9817-PA	6	-	-	-
CG2202	CG2202-PA	14	-	-	ZAD
CG11122	CG11122-PA	5	-	-	-
CG11105	CG11105-PA	1	-	-	efhand (x2)
CG11105	CG11105-PB	1	-	-	efhand (x2)
CG11696	CG11696-PA	9	-	-	AT_hook, ZAD
CG11695	CG11695-PA	9	-	-	ZAD
CG15725	CG15725-PA	2	-	-	BTB
CG4318	CG4318-PA	3	-	-	ZAD
CG11071	CG11071-PA	6	-	-	-
CG32611	CG32611-PB	4	-	-	BTB
CG9203	CG9203-PA	1	x	x	-
CG9215	CG9215-PA	5	-	-	ZAD
CG8944	CG8944-PB	8	-	-	-
CG8944	CG8944-PA	1	x	-	-
disco-r	CG32577-r-PA	4	-	-	-
disco	CG9908-PA	2	-	x	-
hang	CG32575-PA	19	-	-	ZAD
hang	CG32575-PB	19	-	-	ZAD
CG9609	CG9609-PA	9	-	x	-

Table A.1 *continued on next page*

Table A.1 *continued from previous page*

Symbol	Protein	# ZnF	conserved in		additional domains
			C	M	
CG13005	CG13005-PA	1	-	-	-
CG6769	CG6769-PA	3	x	x	-
CG6470	CG6470-PA	2	-	-	-
CG7101	CG7101-PA	9	-	-	-
CG14200	CG14200-PA	1	-	-	PWWP
CG12701	CG12701-PA	6	-	-	-
CG1314	CG1314-PA	1	-	-	-
CG1529	CG1529-PA	5	-	-	ZAD
Chromosomearm 2L					
CG4133	CG4133-PA	1	-	-	-
EP2237	CG4427-PA	3	-	x	-
ush	CG2762-PA	9	-	-	-
CG4887	CG4887-PA	1	x	x	G-patch, RRM_1, zf-RanBP
CG4896	CG4896-PC	1	-	-	G-patch, RRM_1, zf-RanBP
CG4896	CG4896-PD	1	-	-	G-patch, RRM_1, zf-RanBP
CG4896	CG4896-PA	1	-	-	G-patch, zf-RanBP
CG4896	CG4896-PB	1	-	-	G-patch, zf-RanBP
CG31666	CG31666-PA	2	-	-	BTB
CG31666	CG31666-PB	1	-	-	BTB
CG31670	CG31670-PA	6	x	x	-
CG9866	CG9866-PA	1	-	-	-
CG18555	CG18555-PA	6	-	-	ZAD
CG3485	CG3485-PA	4	-	-	ZAD
drm	CG10016-PA	2	-	-	-
sob	CG3242-PA	5	-	x	-
odd	CG3851-PA	4	x	-	-
bowl	CG10021-PA	5	-	-	-
CG3407	CG3407-PA	7	-	-	-
CG17612	CG17612-PA	10	-	-	ZAD
CG15435	CG15435-PA	2	-	-	ZAD
CG15436	CG15436-PA	8	-	-	ZAD
Cf2	CG11924-PD	3	-	-	-

Table A.1 *continued on next page*

Table A.1 *continued from previous page*

Symbol	Protein	# ZnF	conserved in		additional
			C	M	domains
Cf2	CG11924-PA	7	-	-	-
Cf2	CG11924-PB	6	-	-	-
l(2)05714	CG8886-PB	1	-	-	-
Kr-h1	CG18783-h1-PB	8	-	-	-
Kr-h1	CG18783-h1-PA	8	-	-	-
CG31642	CG31642-PA	1	-	-	ZZ
CG31632	CG31632-PA	6	x	x	-
CG4496	CG4496-PA	6	-	-	-
chm	CG5229-PA	1	-	x	MOZ/SAS, zf-C2HC
fu2	CG9233-PA	9	-	-	ZAD
scat	CG3766-PA	1	x	x	-
zf30C	CG3998-PA	13	-	-	AT_hook
CG13123	CG13123-PA	3	-	-	ZAD
CG13131	CG13131-PA	1	-	-	-
CG5694	CG5694-PA	1	-	-	-
CG12299	CG12299-PA	11	-	-	-
ab	CG4807-PA	2	-	-	BTB
ab	CG4807-PB	2	-	-	BTB
CG32830	CG32830-PA	2	-	-	-
salr	CG4881-PA	8	-	x	-
salm	CG6464-PA	7	x	x	-
crol	CG14938-PA	18	-	x	-
crol	CG14938-PB	16	-	x	-
crol	CG14938-PC	12	-	x	-
crol	CG14938-PD	15	-	x	-
CG6792	CG6792-PA	7	-	-	BTB
CG9932	CG9932-PA	16	-	-	-
CG5204	CG5204-PA	8	-	-	-
CG15482	CG15482-PA	1	-	-	-
CG15286	CG15286-PA	1	-	-	ZZ
elB	CG4220-PB	1	-	-	-
elB	CG4220-PA	1	-	-	-
elB	CG4220-PC	1	-	-	-
noc	CG4491-PA	1	-	x	-
CG31835	CG31835-PA	1	-	-	ZZ
CG15269	CG15269-PA	8	-	-	-

Table A.1 *continued on next page*

Table A.1 *continued from previous page*

Symbol	Protein	# ZnF	conserved in		additional domains
			C	M	
esg	CG3758-PA	5	x	x	-
wor	CG4158-PA	6	-	-	-
sna	CG3956-PA	5	-	-	-
l(2)35Ea	CG4148-PA	6	-	-	AT_hook, ZAD
CG17328	CG17328-PA	6	-	-	ZAD
her	CG4694-PA	4	-	-	-
CG31782	CG31782-PA	12	-	-	-
CG31782	CG31782-PC	8	-	-	ZAD
CG31782	CG31782-PB	12	-	-	-
CG31782	CG31782-PD	3	-	-	ZAD
CG17912	CG17912-PA	2	x	x	-
CG10348	CG10348-PA	3	x	-	-
CG31753	CG31753-PA	9	-	x	-
CG10431	CG10431-PA	6	-	-	ZAD
CG17568	CG17568-PA	6	-	-	ZAD
CG10137	CG10137-PA	1	x	x	UVR, zf-TRAF
CG10263	CG10263-PD	1	-	x	-
CG10263	CG10263-PA	1	-	x	-
CG10263	CG10263-PC	1	-	x	-
CG10366	CG10366-PA	7	-	-	ZAD
CG10462	CG10462-PA	8	-	-	-
CG10631	CG10631-PA	4	-	-	THAP (x26)
CG31612	CG31612-PA	16	-	-	-
tsh	CG1374-PA	3	-	-	-
tsh	CG1374-PB	3	-	-	-
tiptop	CG12630-PA	5	-	x	-
CG1832	CG1832-PA	7	-	-	-
Chromosomearm 2R					
CG30431	CG30431-PA	6	-	-	ZAD
jing	CG9403-PD	3	-	x	-
jing	CG9403-PA	3	-	x	-
CG3274	CG3274-PA	1	x	x	ARID
CG30443	CG30443-PA	9	-	-	-
CG12842	CG12842-PA	1	-	-	-
CG1845	CG1845-PA	1	x	x	BROMO, PHD, PWWP
az2	CG1605-PA	8	-	-	-

Table A.1 *continued on next page*

Table A.1 *continued from previous page*

Symbol	Protein	# ZnF	conserved in		additional
			C	M	domains
CG1603	CG1603-PA	7	-	-	-
CG1602	CG1602-PA	8	-	-	-
CG12769	CG12769-PA	7	x	-	-
CG12769	CG12769-PB	7	x	-	-
rgr	CG8643-PA	8	-	-	-
l(2)k10201	CG13951-PA	2	-	x	-
CG1663	CG1663-PA	6	-	-	-
CG18446	CG18446-PA	3	-	-	-
CG12744	CG12744-PA	3	-	-	-
CG18011	CG18011-PA	19	-	-	ZAD
CG12909	CG12909-PA	1	x	x	-
lola	CG12052-PG	2	-	-	BTB
lola	CG12052-PX	1	-	-	BTB
lola	CG12052-PN	3	-	-	BTB
lola	CG12052-PY	2	-	-	BTB
lola	CG12052-PW	2	-	-	BTB
lola	CG12052-PO	2	-	-	BTB
lola	CG12052-PJ	2	-	-	BTB
lola	CG12052-PP	2	-	-	BTB
lola	CG12052-PC	2	-	-	BTB
lola	CG12052-PI	2	-	-	BTB
lola	CG12052-PR	2	-	-	BTB
lola	CG12052-PT	2	-	-	BTB
lola	CG12052-PF	2	-	-	BTB
lola	CG12052-PK	2	-	-	BTB
lola	CG12052-PB	2	-	-	BTB
lola	CG12052-PS	1	-	-	BTB
lola	CG12052-PA	2	-	-	BTB
lola	CG12052-PL	2	-	-	BTB
lola	CG12052-PQ	2	-	-	BTB
CG30020	CG30020-PA	15	-	-	ZAD
CG12942	CG12942-PA	10	-	-	ZAD
CG33473	CG33473-PB	3	x	x	-
CG12391	CG12391-PA	4	-	-	AT_hook, ZAD
shn	CG7734-PD	8	x	x	-
shn	CG7734-PA	8	x	x	-
stil	CG8592-PA	1	-	-	-

Table A.1 *continued on next page*

Table A.1 *continued from previous page*

Symbol	Protein	# ZnF	conserved in		additional
			C	M	domains
sug	CG3850-PA	5	-	x	-
CG6701	CG6701-PA	1	-	-	-
cg	CG8367-PA	11	-	-	-
cg	CG8367-PB	11	-	-	-
cg	CG8367-PC	11	-	-	-
CG17385	CG17385-PA	7	-	-	-
CG17390	CG17390-PA	20	-	x	-
CG12863	CG12863-PA	1	x	x	zf-CCHC (x2)
CG10265	CG10265-PA	1	-	-	-
CG11798	CG11798-PA	5	-	x	-
CG8089	CG8089-PA	7	-	-	-
CG8092	CG8092-PA	5	-	x	AT_hook (x3)
CG8388	CG8388-PA	9	-	-	AT_hook, ZAD
CG30096	CG30096-PA	1	-	-	-
CG4282	CG4282-PA	10	-	-	AT_hook (x2), ZAD
CG15710	CG15710-PA	4	-	-	GATA
tef	CG8961-PA	3	-	-	-
CG30460	CG30460-PA	1	-	-	-
CG30460	CG30460-PB	1	-	-	-
CG30460	CG30460-PC	2	-	-	-
MESR4	CG4903-PA	8	-	-	PHD
sbb	CG5580-PA	1	-	x	-
CG15073	CG15073-PA	9	-	-	ZAD
CG11906	CG11906-PA	7	-	-	-
CG10543	CG10543-PA	9	-	-	-
CG10543	CG10543-PC	9	-	-	-
CG10543	CG10543-PB	8	-	-	-
grau	CG33133-PA	8	-	-	ZAD
CG10321	CG10321-PA	5	-	-	ZAD
px	CG4444-PA	1	-	-	-
CG9895	CG9895-PA	3	-	-	-
CG9890	CG9890-PA	3	x	x	-
pita	CG3941-PA	10	-	-	ZAD
pita	CG3941-PB	10	-	-	-
ken	CG5575-PA	3	-	-	BTB
CG3065	CG3065-PA	5	-	-	-

Table A.1 *continued on next page*

Table A.1 *continued from previous page*

Symbol	Protein	# ZnF	conserved in		additional
			C	M	domains
CG3065	CG3065-PC	5	-	-	-
enok	CG11290-PA	1	x	x	MOZ/SAS, PHD
CG11414	CG11414-PA	4	x	x	zf-C3HC4
CG4707	CG4707-PA	10	-	-	ZAD
key	CG16910-PA	1	-	-	-
CG2790	CG2790-PA	2	x	x	DnaJ
Kr	CG3340-PA	5	-	-	-
Chromosomearm 3L					
CG1231	CG1231-PA	3	-	x	zF-U1
CG1233	CG1233-PB	10	-	-	-
CG1233	CG1233-PA	10	-	-	-
CG17181	CG17181-PA	5	-	-	-
nerfin-1	CG13906-1-PA	3	-	x	-
CG2199	CG2199-PA	7	-	-	-
CG2199	CG2199-PB	7	-	-	-
CG1244	CG1244-PA	7	x	-	-
CG32486	CG32486-PD	1	-	x	zf-TRAF
CG14962	CG14962-PA	4	x	x	-
CG12029	CG12029-PA	3	x	x	-
CG12605	CG12605-PA	5	x	-	-
CG12605	CG12605-PC	5	x	-	-
scrt	CG1130-PA	5	-	x	-
CG11586	CG11586-PA	1	x	x	-
CG4603	CG4603-PA	1	-	x	Otu, ubiquitin
CG5249	CG5249-PA	5	x	x	SET
CG13287	CG13287-PA	2	x	x	-
CG13296	CG13296-PA	4	-	x	-
CG10274	CG10274-PA	9	-	-	ZAD
D19B	CG10270-PA	12	-	-	ZAD
D19A	CG10269-PA	12	-	-	AT_hook, ZAD
CG7386	CG7386-PA	12	-	-	ZAD
CG10147	CG10147-PA	9	-	-	ZAD
CTCF	CG8591-PA	11	-	x	AT_hook
CG8209	CG8209-PA	1	x	x	UBA,UBX
CG6765	CG6765-PA	2	-	-	BTB
MTF-1	CG3743-1-PA	6	-	x	-
MTF-1	CG3743-1-PB	2	-	x	-

Table A.1 *continued on next page*

Table A.1 *continued from previous page*

Symbol	Protein	# ZnF	conserved in		additional
			C	M	domains
phol	CG3445-PA	4	-	-	-
CG8108	CG8108-PA	2	-	x	-
klu	CG12296-PA	4	x	-	-
CG7512	CG7512-PA	2	x	x	-
CG7368	CG7368-PA	4	-	-	-
CG10654	CG10654-PA	6	-	-	ZAD
CG10754	CG10754-PA	1	x	x	-
CG11008	CG11008-PA	1	-	-	LRV
CG14117	CG14117-PA	1	-	-	-
sens	CG32120-PA	4	-	-	-
CG32121	CG32121-PA	2	-	-	BTB
Meics	CG8474-PA	12	-	-	ZAD
CG17361	CG17361-PA	1	-	-	ZAD
CG17359	CG17359-PA	5	-	-	ZAD
Trl	CG33261-PA	1	-	-	BTB
Trl	CG33261-PB	1	-	-	BTB
CG9425	CG9425-PA	1	-	x	TPR, zf-CCCH
CG9425	CG9425-PB	1	-	x	TPR, zf-CCCH
CG7372	CG7372-PA	9	-	-	-
CkIIalpha-i1	CG6215-i1-PA	4	-	-	-
CG18081	CG18081-PA	1	-	-	-
CG15715	CG15715-PA	1	x	x	-
Zn72D	CG5215-PB	3	x	x	DZF
Zn72D	CG5215-PA	3	x	x	-
CG18265	CG18265-PA	10	x	-	-
Pep	CG6143-PA	4	-	x	-
Pep	CG6143-PB	4	-	x	-
Pep	CG6143-PC	4	-	x	-
CG7271	CG7271-PA	1	-	-	-
term	CG4216-PA	1	-	-	-
CG6885	CG6885-PA	1	-	-	-
Su(z)12	CG8013-PB	1	-	x	-
Su(z)12	CG8013-PA	1	-	x	-
kin17	CG5649-PA	1	x	x	KOW
CG11456	CG11456-PA	8	-	-	-
CG7752	CG7752-PA	6	-	-	-
Aef1	CG5683-PA	4	-	-	-

Table A.1 *continued on next page*

Table A.1 *continued from previous page*

Symbol	Protein	# ZnF	conserved in C	conserved in M	additional domains
Neu2	CG7204-PA	8	-	-	ZAD
CG11247	CG11247-PA	11	-	-	-
CG14451	CG14451-PA	3	-	-	TNFR_c6
jim	CG11352-PB	9	-	-	-
CG10712	CG10712-PA	1	-	-	CHROMO
Ssl1	CG11115-PA	1	x	x	Ssl1
Chromosomearm 3R					
hkb	CG9768-PA	3	-	-	-
CG14655	CG14655-PA	8	-	-	-
opa	CG1133-PA	4	x	x	-
CG14667	CG14667-PA	2	-	-	ZAD
noi	CG2925-PA	1	x	x	-
MTA1-like	CG2244-like-PA	1	x	x	BAH, ELM2, MYB, GATA
MTA1-like	CG2244-like-PB	1	x	x	BAH, ELM2, MYB, GATA
cas	CG2102-PA	4	-	x	AT_hook
CG10979	CG10979-PA	6	-	x	-
CG10999	CG10999-PA	2	x	x	-
CG10999	CG10999-PC	2	x	x	-
CG1024	CG1024-PA	5	-	-	AT_hook
CG10267	CG10267-PA	5	-	-	ZAD
rn	CG32466-PA	6	x	x	-
rn	CG32466-PB	5	x	x	-
CG2678	CG2678-PA	7	-	-	ZAD
CG7963	CG7963-PA	5	-	-	ZAD
hb	CG9786-PA	6	x	-	-
CG8145	CG8145-PA	5	-	-	ZAD
CG11762	CG11762-PA	5	-	-	ZAD
CG8159	CG8159-PA	5	-	-	ZAD
CG9793	CG9793-PA	5	-	-	ZAD
CG9797	CG9797-PA	5	-	-	ZAD
CG11966	CG11966-PA	2	-	-	-
CG11971	CG11971-PA	6	-	-	ZAD
CG11984	CG11984-PA	1	x	x	ZZ
CG16779	CG16779-PA	6	-	x	ELM2, MYB
CG8301	CG8301-PA	13	-	-	ZAD

Table A.1 *continued on next page*

Table A.1 *continued from previous page*

Symbol	Protein	# ZnF	conserved in		additional domains
			C	M	
CG8319	CG8319-PA	8	-	-	ZAD
CG16899	CG16899-PA	1	x	x	FH
CG8478	CG8478-PB	3	-	-	-
CG8478	CG8478-PA	3	-	-	-
CG8484	CG8484-PA	11	-	-	-
nerfin-2	CG12809-2-PA	3	x	-	-
CG6254	CG6254-PA	8	-	-	ZAD
CG6325	CG6325-PA	1	-	x	-
CG4820	CG4820-PA	4	-	-	ZAD
CG6689	CG6689-PA	7	-	-	THAP, ZAD
CG31441	CG31441-PA	5	-	-	ZAD
CG31388	CG31388-PA	8	-	-	ZAD
CG6791	CG6791-PA	15	-	-	
CG6791	CG6791-PB	17	-	-	
CG14710	CG14710-PA	5	-	-	ZAD
CG6808	CG6808-PA	5	-	-	ZAD
CG14711	CG14711-PA	5	-	-	AT_hook, ZAD
CG6813	CG6813-PA	2	-	-	ZAD
CG18764	CG18764-PA	5	-	-	ZAD
CG18476	CG18476-PA	17	-	-	ZAD
CG6930	CG6930-PA	4	-	-	-
CG3281	CG3281-PA	8	-	-	ZAD
MBD-R2	CG10042-R2-PA	1	-	x	MBD, PHD
MBD-R2	CG10042-R2-PB	1	-	x	MBD, PHD, THAP
CG5245	CG5245-PA	15	x	-	-
CG17319	CG17319-PA	1	-	-	LRR (x3)
trx	CG8651-PA	1	-	x	FYRC, FYRN, PHD (x2), SET
trx	CG8651-PB	1	-	x	FYRC, FYRN, PHD (x2), SET
su(Hw)	CG8573-PA	12	-	-	-
CG7987	CG7987-PA	15	-	-	-
CG6654	CG6654-PA	10	-	-	ZAD
Cp190	CG6384-PA	3	-	-	BTB
CG31392	CG31392-PA	6	-	-	-

Table A.1 *continued on next page*

Table A.1 *continued from previous page*

Symbol	Protein	# ZnF	conserved in		additional
			C	M	domains
spn-E	CG3158-E-PA	1	-	-	HA2, Helicase_C, TUDOR
CG10309	CG10309-PA	4	-	-	ZAD
CG17803	CG17803-PA	7	-	-	ZAD
CG17806	CG17806-PA	5	-	-	ZAD
CG17802	CG17802-PA	5	-	-	ZAD
CG17801	CG17801-PA	4	-	-	ZAD
CG7357	CG7357-PA	5	-	-	ZAD
sr	CG7847-PA	3	x	x	-
sr	CG7847-PB	3	x	x	-
CG7985	CG7985-PA	1	x	x	-
gl	CG7672-PA	5	x	-	-
gl	CG7672-PB	4	x	-	-
fru	CG14307-PB	1	-	-	BTB
fru	CG14307-PF	1	-	-	BTB
fru	CG14307-PG	2	-	-	BTB
fru	CG14307-PH	2	-	-	BTB
fru	CG14307-PK	2	-	-	BTB
fru	CG14307-PC	2	-	-	BTB
fru	CG14307-PE	2	-	-	BTB
fru	CG14307-PI	2	-	-	BTB
fru	CG14307-PJ	2	-	-	BTB
CG7691	CG7691-PA	3	-	-	-
CG31224	CG31224-PA	14	-	-	-
sqz	CG5557-PA	5	-	-	-
CG5316	CG5316-PB	2	-	x	HIT,RHS
CG17186	CG17186-PA	3	-	-	-
CG4424	CG4424-PA	5	-	-	ZAD
CG4854	CG4854-PA	5	-	-	ZAD
CG4413	CG4413-PA	5	-	-	ZAD
CG4936	CG4936-PA	5	-	-	ZAD
CG4360	CG4360-PA	9	-	-	-
CG4813	CG4813-PA	1	-	x	-
lmd	CG4677-PB	5	-	x	-
lmd	CG4677-PA	5	-	x	-
CG31365	CG31365-PA	6	-	-	ZAD
CG4374	CG4374-PA	2	-	-	Tubulin-binding

Table A.1 *continued on next page*

Table A.1 *continued from previous page*

Symbol	Protein	# ZnF	conserved in		additional domains
			C	M	
CG5669	CG5669-PA	3	x	x	-
CG13617	CG13617-PA	1	-	-	-
CG13620	CG13620-PA	7	-	-	-
CG11375	CG11375-PA	1	x	x	AT_hook, BROMO (x5), BAH (x2), HMG
CG31381	CG31381-PA	1	x	x	IPPT
CG31510	CG31510-PA	2	-	-	-
CG10669	CG10669-PA	17	-	-	ZAD
CG11902	CG11902-PA	24	-	-	ZAD
CG4730	CG4730-PA	6	-	-	ZAD
CG18437	CG18437-PA	1	x	x	-
CG33213	CG33213-PA	5	-	-	-
CG31053	CG31053-PA	1	x	-	-
CG1894	CG1894-PA	1	-	-	MOZ/SAS
wdn	CG1454-PA	8	-	-	-
CG1647	CG1647-PA	4	-	-	ZAD
CG7928	CG7928-PA	7	-	-	ZAD
Sry-beta	CG7938-beta-PA	6	-	-	ZAD
Sry-delta	CG17958-delta-PA	7	-	-	ZAD
zfh1	CG1322-PB	9	x	x	HOX
zfh1	CG1322-PA	7	x	x	HOX
CG12071	CG12071-PB	3	-	-	-
CG12071	CG12071-PA	1	-	-	-
CG11317	CG11317-PA	2	-	-	-
CG12054	CG12054-PA	3	-	x	-
CG12114	CG12114-PA	1	-	-	-
CG1792	CG1792-PA	5	-	-	ZAD
ttk	CG1856-PA	2	-	-	AT_hook, BTB
ttk	CG1856-PC	2	-	-	AT_hook, BTB
Chromosome 4					
ci	CG2125-PA	5	x	x	-
CG2052	CG2052-PA	7	-	-	-
CG2052	CG2052-PB	7	-	-	-
zfh2	CG1449-PA	13	x	x	HOX (x3)
pho	CG17743-PA	4	-	x	-

Table A.2: All *Drosophila melanogaster* ZADs. Genes marked with * do not have ZnF domains; the number in brackets denote the number of orthologs in *D. pseudoobscura* if there is more than one; Dpse, *D. pseudoobscura*; Agam, *A. gambiae*.

CG number	Symbol	Subset 1	Subset 2	Sub-group	Gene cluster	EST	conserved in Dpse	conserved in Agam
				Chromosome X				
CG2711	dwg	x	-	C	1	x	x	-
CG2712		x	-	C	1	x	x	-
CG12219		-	x	a		x	x	-
CG14441		-	x			-	x	-
CG3032		x	-			x	x	-
CG15367	DIP1	x	-			-	x	-
CG2889		-	x	a		x	x	-
CG2202		x	-			x	x	-
CG11696		-	x	a	2	x	x	-
CG11695		-	x	a	2	x	x	-
CG4318		x	-			-	-	-
CG9215		x	-			x	x(2)	-
CG32575		-	x			x	x	-
CG1529		x	-			x	x	-
				Chromosomearm 2L				
CG11371*		-	x			x	x	-
CG18555		x	-	A	3	x	-	-
CG3485		x	-	A	3	-	-	-
CG17612		x	-	A		x	-	-
CG15435		-	x	d	4	x	x	x
CG15436		x	-	A	4	x	x(19)	-
CG9233	fu2	-	x			x	x	-
CG13123		-	x			x	x	-
CG4148	l(2)35Ea	-	x	b		x	x	-
CG17328		-	x			-	x	x
CG31782-PC		-	x	A	5	-	-	-
CG31782-PD		x	-	A	5	-	-	-
CG10431		x	-			x	x	-
CG17568		-	x	b		-	x	-
CG10366		-	x	b		x	x	-
				Chromosomearm 2R				
CG30431		x	-			x	x	-

Table A.2 *continued on next page*

Table A.2 *continued from previous page*

CG number	Symbol	Subset 1	Subset 2	Sub-group	Gene cluster	EST	conserved in Dpse	conserved in Agam
CG18011		x	-			x	x	-
CG30020		x	-		6	x	x	-
CG12942		x	-		6	x	x	-
CG12391		x	-			x	x	-
CG10108*	phyl	-	x			x	x	-
CG8388		-	x			x	x	-
CG4282		x	-			x	x	-
CG15073		-	x	c		x	x	-
CG33133		-	x	c		x	x	-
CG10321		x	-			x	x	x
CG3941	pita	-	x			x	x	x
CG4707		-	x			x	x	-
		Chromosomearm 3L						
CG10274		x	-	B	7	x	x	-
CG10270	D19B	x	-	B	7	x	x	-
CG10269	D19A	x	-	B	7	x	x	-
CG7386		x	-	B	7	x	x	-
CG10147		x	-			-	x(2)	-
CG10654		x	-			x	x	-
CG8474	Meics	x	-			x	x	-
CG17361		x	-		8	x	x(2)	-
CG17359		x	-		8	x	x	-
CG7204		x	-	A		-	-	-
		Chromosomearm 3R						
CG14667		-	x	d		x	x	-
CG10267		-	x			x	x	-
CG2678		x	-			x	x(4)	-
CG7963		x	-			-	x	-
CG8145		-	x	d	9	x	x	-
CG11762		-	x	d	9	-	x	-
CG8159		-	x	d	9	x	x	-
CG9793		-	x	d	9	-	-	-
CG9797		-	x	d	9	x	x	-
CG11971		-	x			x	x	-
CG8301		-	x			x	x	-
CG8319		x	-			x	x	-
CG6254		-	x	b		x	x	-

Table A.2 *continued on next page*

Table A.2 *continued from previous page*

CG number	Symbol	Subset 1	Subset 2	Sub-group	Gene cluster	EST	conserved in Dpse	conserved in Agam
CG4820		x	-		10	x	x	-
CG6689		-	x	d	10	x	x(4)	-
CG31441		-	x		11	x	-	-
CG31388		-	x	d	11	x	-	-
CG14710		-	x		12	x	x	-
CG6808		-	x		12	x	x	-
CG14711		-	x		12	x	x	-
CG6813		-	x	d	12	-	x	-
CG18764		-	x	d	12	x	x(2)	-
CG18476		-	x	e		x	x	-
CG3281		-	x			x	x	-
CG6654		-	x			x	x	-
CG10309		-	x			-	x	-
CG17803		-	x	d	13	-	-	-
CG17806		-	x	d	13	x	x(2)	-
CG17802		-	x	d	13	x	x	-
CG17801		-	x	d	13	-	-	-
CG7357		-	x	d	13	x	x	-
CG4424		-	x		14	x	x	-
CG4854		-	x	d	14	x	x	-
CG4413		-	x	f	14	x	x	-
CG4936		-	x	f		x	x	-
CG31365		x	-		15	x	x	-
CG31457*		x	-		15	x	x	-
CG31109*		-	x			x	x	x
CG10669		-	x	e	16	x	x	-
CG11902		-	x	e	16	x	x	-
CG4730		x	-			x	-	-
CG1647		-	x			x	x	x
CG7928		x	-		17	x	x	-
CG7938	Sry-b	x	-		17	x	x	-
CG17958	Sry-d	x	-		17	x	x	-
CG1792		-	x	d		x	x	-

Table A.3: All *Drosophila pseudoobscura* ZADs. Genes marked with * have only been identified as fragments; # ZnF, number of found ZnF motifs.

ID	Contig	start end	# exons	strand	subset	# ZnF
Dpse_only1	Contig3266	1091789	1	-	1	0
	Contig6542	1092319				
PG10108	Contig3295	385375	3	-	2	0
	Contig8282	389829				
PG10147-1	Contig1770	151715	2	+	1	9
	Contig8226	153353				
PG10147-2	Contig1770	147421	2	+	1	9
	Contig8226	148850				
PG10267	Contig4374	1047897	4	-	2	5
	Contig4847	1049303				
PG10269	Contig5115	197484	1	-	1	12
	Contig7866	199988				
PG10270	Contig3914	410023	1	+	1	12
	Contig1507	412341				
PG10274	Contig3914	417666	3	-	1	12
	Contig1507	420396				
PG10309	Contig3681	1088682	4	-	2	4
	Contig434	1091730				
PG10321	Contig3295	1738356	1	+	1	5
	Contig8282	1740896				
PG10366	Contig903	128453	3	+	2	8
	Contig7450	130340				
PG10431	Contig3635	1050211	1	-	1	6
	Contig7816	1052460				
PG10654	Contig4922	2836152	2	-	1	6
	Contig3818	2838126				
PG10669	Contig4374	1466490	6	+	2	17
	Contig4847	1469684				
PG11371	Contig2076	626914	4	+	2	0
	Contig2470	630720				
PG11695	Contig1045	127928	4	-	2	10
	Contig3832	129665				
PG11696	Contig1045	130429	4	-	2	9
	Contig3832	132593				

Table A.3 *continued on next page*

Table A.3 *continued from previous page*

ID	Contig	start end	# exons	strand	subset	# ZnF
PG11762	Contig3070	215788	3	+	2	5
	Contig6431	216937				
PG11902*	Contig4374	1461028	5	-	2	23
	Contig4847	1465581				
PG11971	Contig3070	390937	3	+	2	5
	Contig6431	393010				
PG12219	Contig4832	781544	4	+	2	4
	Contig7452	783446				
PG12391	Contig6654	558056	1	+	1	4
	Contig5965	559510				
PG12942	Contig2556	549502	6	+	1	10
	Contig5224	551947				
PG13123	Contig2212	207327	3	+	2	3
	Contig7437	208410				
PG14441	Contig2138	40756	7	-	2	3
	Contig8242	51667				
PG14667	Contig3681	227698	2	-	2	3
	Contig434	228506				
PG14710	Contig4374	1540256	4	+	2	5
	Contig4847	1541748				
PG14711	Contig4374	1543669	6	+	2	5
	Contig4847	1545101				
PG15073	Contig3295	1278991	5	+	2	10
	Contig8282	1281128				
PG1529	Contig374	86459	2	+	1	8
	Contig5738	87842				
PG15367	Contig2105	691456	2	+	1	0
	Contig7229	692103				
PG15435	Contig1828	447869	4	-	2	1
	Contig8101	449652				
PG1647	Contig446	721419	4	-	2	3
	Contig3289	725331				
PG17328	Contig2076	76697	2	+	2	6
	Contig2470	78072				
PG17359	Contig2341	303034	1	+	1	7
	Contig3968	304428				

Table A.3 *continued on next page*

Table A.3 *continued from previous page*

ID	Contig	start end	# exons	strand	subset	# ZnF
PG17361-1	Contig2341	305196	1	-	1	1
	Contig3968	305929				
PG17361-2	Contig3681	803854	1	+	1	0
	Contig434	804312				
PG17568	Contig138	595898	2	+	2	7
	Contig706	597567				
PG17802	Contig1193	14179	4	+	2	5
	Contig1472	15422				
PG17806-1	Contig4374	1533692	4	+	2	5
	Contig4847	1535054				
PG17806-2	Contig4374	1531766	4	+	2	4
	Contig4847	1533129				
PG1792	Contig2803	270811	4	+	2	5
	Contig3631	272143				
PG17958	Contig3266	579311	1	-	1	7
	Contig6542	580570				
PG18011	Contig2561	521038	1	-	1	0
	Contig6127	521853				
PG18476	Contig6581	214306	5	-	2	17
	Contig6381	217379				
PG18764-1	Contig4374	1548498	4	+	2	5
	Contig4847	1549888				
PG18764-2	Contig4374	1545578	5	+	2	3
	Contig4847	1546883				
PG2202	Contig4055	547449	4	-	1	13
	Contig7066	550421				
PG2678-1	Contig3070	159400	1	+	1	1
	Contig6431	160329				
PG2678-2	Contig2395	222466	3	-	1	1
	Contig6282	223797				
PG2678-3	Contig917	13987	3	-	1	1
	Contig5291	17735				
PG2678-4	Contig917	162119	3	-	1	0
	Contig5291	163118				
PG2711*	Contig7446	820950	2	+	1	1
	Contig2444	821914				

Table A.3 *continued on next page*

Table A.3 *continued from previous page*

ID	Contig	start end	# exons	strand	subset	# ZnF
PG2712	Contig7446	827092	2	+	1	7
	Contig2444	828407				
PG2889	Contig1277	100709	5	+	2	10
	Contig4006	102775				
PG30020	Contig2556	540619	13	+	1	15
	Contig5224	547169				
PG3032	Contig2105	819461	1	+	1	9
	Contig7229	820789				
PG30431	Contig5727	120204	2	-	1	6
	Contig591	121523				
PG31109*	Contig3553	197	2	-	2	0
	Contig6240	1505				
PG31365	Contig446	472097	3	-	1	6
	Contig3289	474118				
PG31457	Contig446	474650	2	-	1	0
	Contig3289	475797				
PG32575	Contig7446	1903328	8	-	2	18
	Contig2444	1917082				
PG3281	Contig3681	1188025	2	-	2	8
	Contig434	1189770				
PG33133	Contig3479	335691	5	+	2	8
	Contig7346	337722				
PG3941	Contig3342	518095	3	-	2	10
	Contig5418	522789				
PG4148	Contig138	592810	3	+	2	6
	Contig706	594183				
PG4282	Contig3295	819337	3	+	1	10
	Contig8282	821409				
PG4413	Contig534	116524	5	-	2	5
	Contig6747	118152				
PG4424	Contig534	113501	5	-	2	5
	Contig6747	114745				
PG4707	Contig3479	653224	4	-	2	10
	Contig7346	655436				
PG4820	Contig2395	209162	4	+	2	4
	Contig6282	210464				

Table A.3 *continued on next page*

Table A.3 *continued from previous page*

ID	Contig	start end	# exons	strand	subset	# ZnF
PG4854	Contig534	114968	4	+	2	5
	Contig6747	116141				
PG4936	Contig534	118764	5	+	2	5
	Contig6747	120532				
PG6254	Contig152	126626	4	-	2	8
	Contig617	128500				
PG6654	Contig4374	1970757	3	-	2	10
	Contig4847	1972699				
PG6689-1	Contig2395	210497	5	-	2	7
	Contig6282	213122				
PG6689-2	Contig2395	217036	3	-	2	0
	Contig6282	218439				
PG6689-3	Contig2395	218923	3	-	2	0
	Contig6282	220234				
PG6689-4	Contig2395	214479	5	-	2	7
	Contig6282	216649				
PG6808	Contig4374	1541854	4	-	2	5
	Contig4847	1543203				
PG6813	Contig4374	1547422	3	-	2	4
	Contig4847	1548298				
PG7357	Contig1193	12543	4	+	2	5
	Contig1472	13834				
PG7386	Contig5115	200498	2	+	1	12
	Contig7866	202693				
PG7928	Contig3266	586769	1	+	1	7
	Contig6542	588079				
PG7938	Contig3266	583525	2	-	2	6
	Contig6542	584653				
PG7963	Contig3681	1800198	2	+	2	8
	Contig434	1801468				
PG8145	Contig3070	214358	4	+	2	5
	Contig6431	215567				
PG8159	Contig3070	217153	4	+	2	5
	Contig6431	218538				
PG8301	Contig3681	2688822	6	+	2	13
	Contig434	2691824				

Table A.3 *continued on next page*

Table A.3 *continued from previous page*

ID	Contig	start end	# exons	strand	subset	# ZnF
PG8319	Contig3681	36687	2	-	1	8
	Contig434	37957				
PG8388	Contig3479	514918	4	+	2	10
	Contig7346	516847				
PG8474	Contig2999	405944	2	+	1	11
	Contig3887	407600				
PG9215-1	Contig6878	635518	2	+	1	5
	Contig6537	637335				
PG9215-2	Contig6878	631078	2	+	1	0
	Contig6537	632028				
PG9233	Contig4967	1072217	2	-	2	9
	Contig2069	1073998				
PG9797	Contig3070	218967	4	-	2	5
	Contig6431	220449				
SUBA01	Contig903	546730	1	+	1	7
	Contig7450	547656				
SUBA02	Contig903	548732	1	+	1	16
	Contig7450	550561				
SUBA03	Contig903	552024	1	+	1	7
	Contig7450	552977				
SUBA04	Contig903	554135	1	+	1	12
	Contig7450	555646				
SUBA05	Contig5770	149874	1	+	1	16
	Contig924	151634				
SUBA06	Contig5770	146825	1	+	1	7
	Contig924	147790				
SUBA07	Contig5770	144922	1	+	1	10
	Contig924	146118				
SUBA08	Contig5770	138965	1	+	1	10
	Contig924	140311				
SUBA09	Contig1045	135043	1	+	1	6
	Contig3832	135969				
SUBA10*	Contig6573	1368259	1	-	1	11
	Contig4	1369539				
SUBA11	Contig1859	2449	1	+	1	9
	Contig703	3687				

Table A.3 *continued on next page*

Table A.3 *continued from previous page*

ID	Contig	start end	# exons	strand	subset	# ZnF
SUBA12	Contig6573	1360600	1	-	1	8
	Contig4	1361724				
SUBA13*	Contig6573	1365810	1	-	1	9
	Contig4	1366967				
SUBA14	Contig5770	153830	1	+	1	14
	Contig924	155662				
SUBA15	Contig6573	1372209	1	-	1	15
	Contig4	1373909				
SUBA16	Contig6573	1370583	1	-	1	5
	Contig4	1371377				
SUBA17	Contig1763	44148	1	+	1	6
	Contig1932	45074				
SUBA18	Contig1763	153437	1	+	1	11
	Contig1932	154732				
SUBA19	Contig2659	155369	6	-	1	5
	Contig715	157999				

Table A.4: All *Anopheles gambiae* ZADs. Genes marked with * have only been identified as fragments; # ZnF, number of found ZnF motifs.

ID	Contig	start end	# exons	strand	subset	# ZnF
001_Agam	AAAB01003342.1	318 2028	4	+	2	6
003_Agam	AAAB01008978.1	242531 244119	3	-	2	6
004_Agam	AAAB01008978.1	218270 219608	2	+	2	5
005_Agam	AAAB01008859.1	5702702 5704964	3	-	2	10
006_Agam	AAAB01008966.1	1795166 1796307	2	+	2	3
007_Agam	AAAB01008966.1	1792623 1793462	2	+	2	3
008_Agam	AAAB01008966.1	1796722 1798291	4	+	2	7
009_Agam	AAAB01008966.1	1810359 1812086	4	+	2	6
010_Agam	AAAB01008966.1	1788059 1788955	2	+	2	2
011_Agam	AAAB01008966.1	1789676 1791023	4	+	2	7
012_Agam	AAAB01008966.1	1833595 1834715	2	+	2	3
014_Agam	AAAB01008966.1	1799525 1800682	4	+	2	4
015_Agam	AAAB01008966.1	1838258 1839819	4	+	2	7
016_Agam	AAAB01008966.1	1848714 1852676	8	-	2	6
017_Agam	AAAB01008966.1	1808525 1809596	2	+	2	2
018_Agam	AAAB01008966.1	1823302 1824556	3	+	2	4
019_Agam	AAAB01008839.1	861326 863870	2	+	1	7

Table A.4 *continued on next page*

Table A.4 *continued from previous page*

ID	Contig	start end	# exons	strand	subset	# ZnF
020_Agam	AAAB01008987.1	14154999 14158959	6	+	2	4
022_Agam	AAAB01008823.1	189790 197193	2	-	1	11
023_Agam	AAAB01008823.1	202377 212413	8	+	1	26
024_Agam	AAAB01008823.1	231023 232854	4	+	1	9
025_Agam	AAAB01008823.1	223773 225710	4	+	1	9
026_Agam	AAAB01008823.1	234145 235966	4	+	1	9
027_Agam	AAAB01008823.1	221036 223214	4	+	1	9
029_Agam	AAAB01008823.1	227463 229435	4	+	1	9
031_Agam	AAAB01008816.1	711337 714968	4	-	2	6
032_Agam	AAAB01008807.1	9260878 9262796	2	+	2	0
033_Agam	AAAB01008807.1	9268498 9270947	2	+	2	0
034_Agam	AAAB01008807.1	9272078 9274488	2	+	2	0
035_Agam	AAAB01008807.1	9263869 9266335	3	+	2	0
036_Agam	AAAB01008807.1	9281817 9285709	7	+	2	2
037_Agam	AAAB01008807.1	9276007 9280217	4	+	2	0
039_Agam	AAAB01008807.1	2251513 2253126	3	+	2	9
040_Agam	AAAB01008807.1	2245368 2248934	5	+	1	17
042_Agam	AAAB01008984.1	7392417 7394377	3	-	2	6

Table A.4 *continued on next page*

Table A.4 *continued from previous page*

ID	Contig	start end	# exons	strand	subset	# ZnF
043_Agam	AAAB01008984.1	7651073 7652549	4	+	2	6
044_Agam	AAAB01008898.1	36505 41858	7	+	1	8
045_Agam	AAAB01008879.1	2488537 2490247	3	-	2	8
046_Agam	AAAB01008980.1	10750676 10751842	1	-	1	0
047_Agam	AAAB01008968.1	1183797 1185810	6	+	2	3
048_Agam	AAAB01008964.1	9100893 9103129	2	-	1	5
049_Agam	AAAB01008964.1	8955045 8956457	3	+	2	9
050_Agam	AAAB01008964.1	8960957 8962349	3	+	2	9
051_Agam*	AAAB01008964.1	8958815 8960354	3	+	2	10
052_Agam	AAAB01008964.1	8956763 8958144	3	+	2	9
053_Agam	AAAB01008964.1	8962687 8964176	3	+	2	9
054_Agam	AAAB01008848.1	615815 617426	3	-	2	7
055_Agam	AAAB01008846.1	6186429 6188449	4	-	2	6
056_Agam	AAAB01008845.1	26869 28160	2	-	2	8
057_Agam	AAAB01001316.1	101 1153	2	+	2	3
058_Agam	AAAB01008987.1	8942231 8946260	5	-	1	0
060_Agam	AAAB01008810.1	55492 59925	4	-	2	5
065_Agam	AAAB01008904.1	1476717 1478780	4	-	1	5

Table A.4 *continued on next page*

Table A.4 *continued from previous page*

ID	Contig	start end	# exons	strand	subset	# ZnF
067_Agam	AAAB01008898.1	3902319 3915335	8	+	2	10
070_Agam	AAAB01008880.1	3840486 3842036	2	-	1	9
071_Agam	AAAB01008880.1	3913424 3915681	5	+	1	9
072_Agam	AAAB01008880.1	3843739 3845030	2	+	2	8
073_Agam	AAAB01008859.1	12303141 12304145	1	+	1	0
074_Agam	AAAB01008966.1	427126 430084	5	+	2	10
075_Agam	AAAB01008966.1	1531219 1532483	3	+	2	3
076_Agam	AAAB01008966.1	1536611 1538170	4	+	2	7
077_Agam	AAAB01008966.1	1539606 1541515	5	+	2	10
078_Agam	AAAB01008966.1	1533512 1536084	5	+	2	4
079_Agam	AAAB01008966.1	1526325 1527851	4	+	2	7
080_Agam	AAAB01008964.1	4851982 4854104	2	-	2	13
081_Agam	AAAB01008835.1	420986 425851	5	+	1	16
082_Agam	AAAB01008835.1	1152625 1154585	2	+	2	10
083_Agam	AAAB01008987.1	12031244 12032009	2	-	2	0
084_Agam	AAAB01008817.1	1274424 1277591	2	+	1	0
085_Agam	AAAB01008817.1	1523621 1525127	2	-	1	5
086_Agam	AAAB01008817.1	1527973 1529875	3	-	2	10

Table A.4 *continued on next page*

Table A.4 *continued from previous page*

ID	Contig	start end	# exons	strand	subset	# ZnF
087_Agam	AAAB01008817.1	1271902 1273786	2	-	1	7
090_Agam	AAAB01008986.1	9631769 9633203	2	+	2	2
091_Agam	AAAB01008986.1	10703114 10704470	3	-	2	6
095_Agam	AAAB01008944.1	5884187 5886229	1	+	1	12
096_Agam	AAAB01008944.1	5886659 5888887	1	+	1	12
097_Agam	AAAB01008944.1	5880217 5882352	3	+	2	9
098_Agam	AAAB01008933.1	97870 99724	3	+	2	8
099_Agam	AAAB01008933.1	95027 97124	2	-	2	10
100_Agam	AAAB01008984.1	6616361 6618526	1	+	1	14
101_Agam*	AAAB01008984.1	6789572 6791135	2	+	2	6
102_Agam	AAAB01008984.1	6619276 6621248	2	-	2	11
103_Agam*	AAAB01004242.1	791 1750	2	+	2	1
106_Agam*	AAAB01008978.1	967185 968474	1	-	1	3
107_Agam	AAAB01008978.1	969415 972958	3	-	2	4
108_Agam	AAAB01008978.1	560730 561868	3	+	2	4
109_Agam	AAAB01008978.1	555551 557520	3	+	2	1
110_Agam	AAAB01008968.1	874980 876485	1	+	1	9
111_Agam*	AAAB01008851.1	1345129 1348482	1	-	1	18

Table A.4 *continued on next page*

Table A.4 *continued from previous page*

ID	Contig	start end	# exons	strand	subset	# ZnF
112_Agam*	AAAB01008851.1	1379972 1381126	1	-	1	5
115_Agam	AAAB01008987.1	15667499 15669728	4	-	2	11
116_Agam	AAAB01008964.1	1789497 1791003	2	-	2	9
117_Agam	AAAB01008987.1	7284653 7285978	2	-	1	3
118_Agam	AAAB01008807.1	11616122 11617262	2	+	2	0
120_Agam	AAAB01008960.1	2135940 2137823	2	+	2	10
121_Agam	AAAB01008960.1	2142066 2143802	2	-	2	10
124_Agam	AAAB01008986.1	5290245 5292056	3	-	1	9
125_Agam	AAAB01008986.1	4206238 4207633	3	-	2	6
127_Agam	AAAB01008984.1	595920 599598	4	-	2	10
128_Agam	AAAB01008984.1	700291 702206	6	+	2	5
130_Agam	AAAB01008880.1	1097174 1099079	2	-	2	0
131_Agam	AAAB01008980.1	1495177 1509733	7	+	2	0
132_Agam	AAAB01008859.1	1925316 1926473	2	-	2	5
133_Agam	AAAB01008964.1	11143644 11145279	3	+	2	9
135_Agam	AAAB01008834.1	2156171 2158250	2	-	2	11
136_Agam	AAAB01008835.1	370095 372361	3	+	2	12
138_Agam	AAAB01008952.1	404288 405664	2	+	1	5

Table A.4 *continued on next page*

Table A.4 *continued from previous page*

ID	Contig	start end	# exons	strand	subset	# ZnF
139_Agam	AAAB01008986.1	7357594 7358603	2	-	2	5
140_Agam*	AAAB01008986.1	8974367 8976187	2	-	2	4
141_Agam	AAAB01008986.1	8994702 8996036	2	+	2	4
142_Agam	AAAB01008986.1	9001409 9005816	4	+	2	6
145_Agam	AAAB01008944.1	3989800 3990874	3	-	2	6
147_Agam	AAAB01008905.1	883208 884734	1	+	1	2
148_Agam	AAAB01008984.1	3904514 3908241	8	-	2	0
150_Agam	AAAB01008966.1	1801768 1805085	6	+	2	10
151_Agam	AAAB01008966.1	1826475 1827989	4	+	2	6
152_Agam	AAAB01008966.1	1829677 1830789	3	+	2	3
153_Agam	AAAB01008966.1	1818275 1819818	4	+	2	6
154_Agam	AAAB01008966.1	1835276 1838076	5	+	2	6
155_Agam	AAAB01008966.1	1843162 1845516	4	+	2	2
156_Agam	AAAB01008966.1	1845944 1847303	3	+	2	3
157_Agam	AAAB01008966.1	1852738 1855057	4	-	2	4
158_Agam	AAAB01008807.1	9258842 9260225	2	-	2	5
159_Agam	AAAB01008807.1	2249106 2250815	3	+	2	10
160_Agam	AAAB01008966.1	1528995 1530669	4	+	2	7

Table A.4 *continued on next page*

Table A.4 *continued from previous page*

ID	Contig	start end	# exons	strand	subset	# ZnF
161_Agam	AAAB01008966.1	1519083 1520981	4	+	2	8
162_Agam	AAAB01008966.1	1522563 1524310	4	+	2	7
163_Agam	AAAB01008964.1	4849439 4851399	2	-	2	10
164_Agam	AAAB01008960.1	2138189 2139741	2	+	2	9
165_Agam	AAAB01008960.1	2140288 2141466	2	+	2	4
166_Agam	AAAB01008960.1	2134286 2135817	2	+	2	9
167_Agam	AAAB01008960.1	2123749 2125370	2	+	2	9
168_Agam	AAAB01008980.1	1489766 1490813	2	+	2	3
169_Agam	AAAB01008986.1	8987211 8988961	3	-	2	1

Table A.5: Perl scripts used in this study. All scripts can be found on the accompanying CD.

Name	short description
`parseHMMER.pl`	parses the output of a `hmmsearch` run and extracts the sequences
`parsePfam.pl`	parses the pfam output and annotates the number of C2H2 zinc-finger domain hits and hits to additional domains (+ their e-value)
`fractionate.pl`	fractionates large genomic contigs into 2,000 bp sequences that overlap by 200 bp
`divide.pl`	divides the sequences into 1,000 sequences each per file
`RunGenScan.pl`	runs `GenScan` on the input (DNA) sequence and returns an EMBL format entry for the coding sequence
`ExtractGI.pl`	extracts the GI number from the `estwisedb` output
`GetOrganism.pl`	extracts the species name and count the number of hits to ESTs of these species
`formatDB.pl`	formats the trace DB for BLAST
`trace2ZAD.pl`	searches in unassembled trace sequences for ZAD coding sequences
`getbestZAD.pl`	gets the best hit to a D. melanogaster ZAD
`Orthology.pl`	performs the reciprocal blast procedure in order to identifiy putative orthologous genes
`dme2genome.pl`	searches for orthologous genes in genome sequences and extracts the (similar) coding-sequences
`runPAML.pl`	runs CodonML with the specified parameters on a large number of genes
`formatAlignment.pl`	maps the amino acid alignment to the DNA coding sequences
`ParsePAML.pl`	extracts information from the PAML output
`runPHD.pl`	runs PHD on a multiple sequence alignment by putting each sequence at the top of the alignment
`consensusPHD.pl`	calculates the consensus secondary structure prediction

Appendix B

Figures

B.1 Alignments

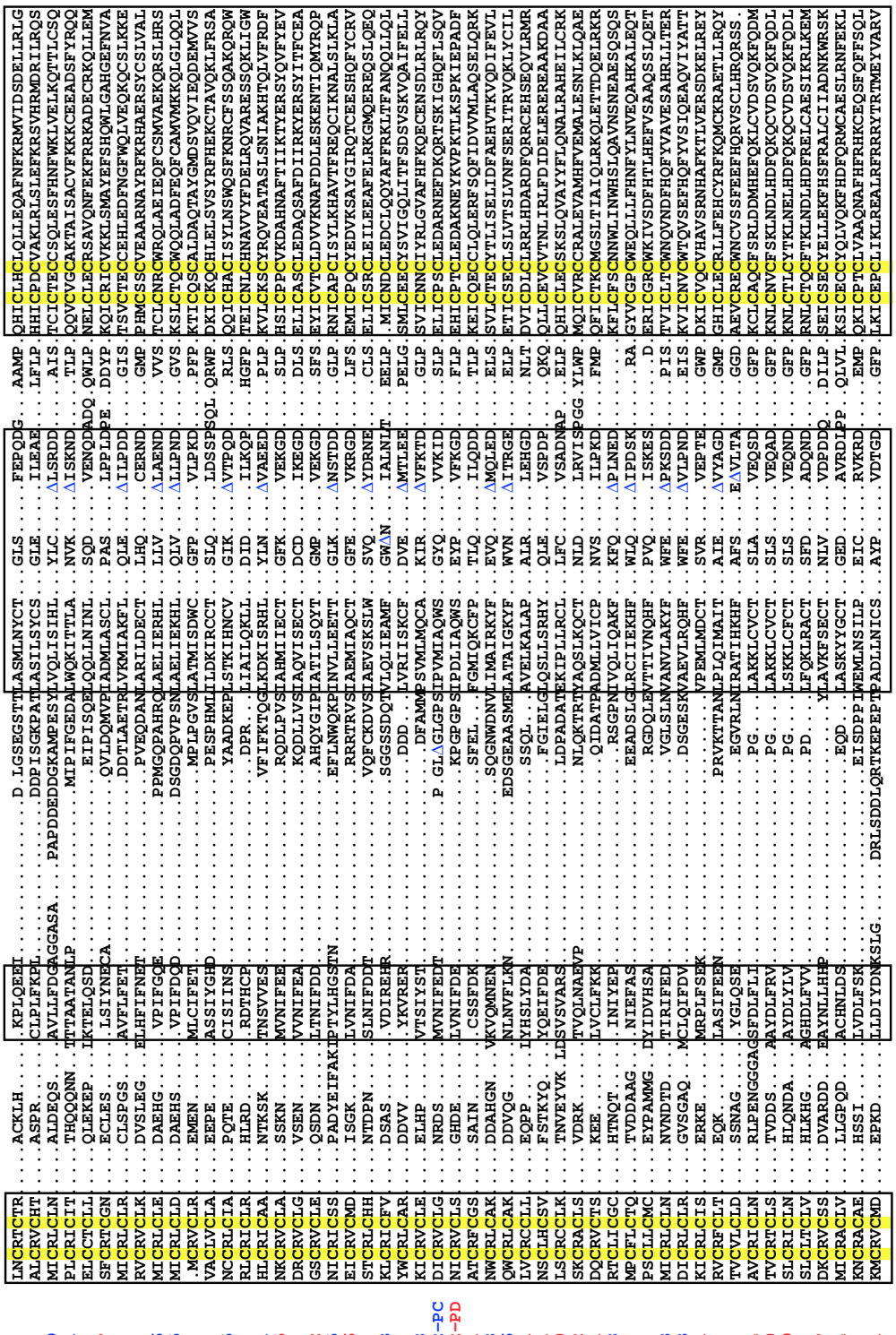


Figure B.1 *continued on next page*

Figure B.1 *continued from previous page*

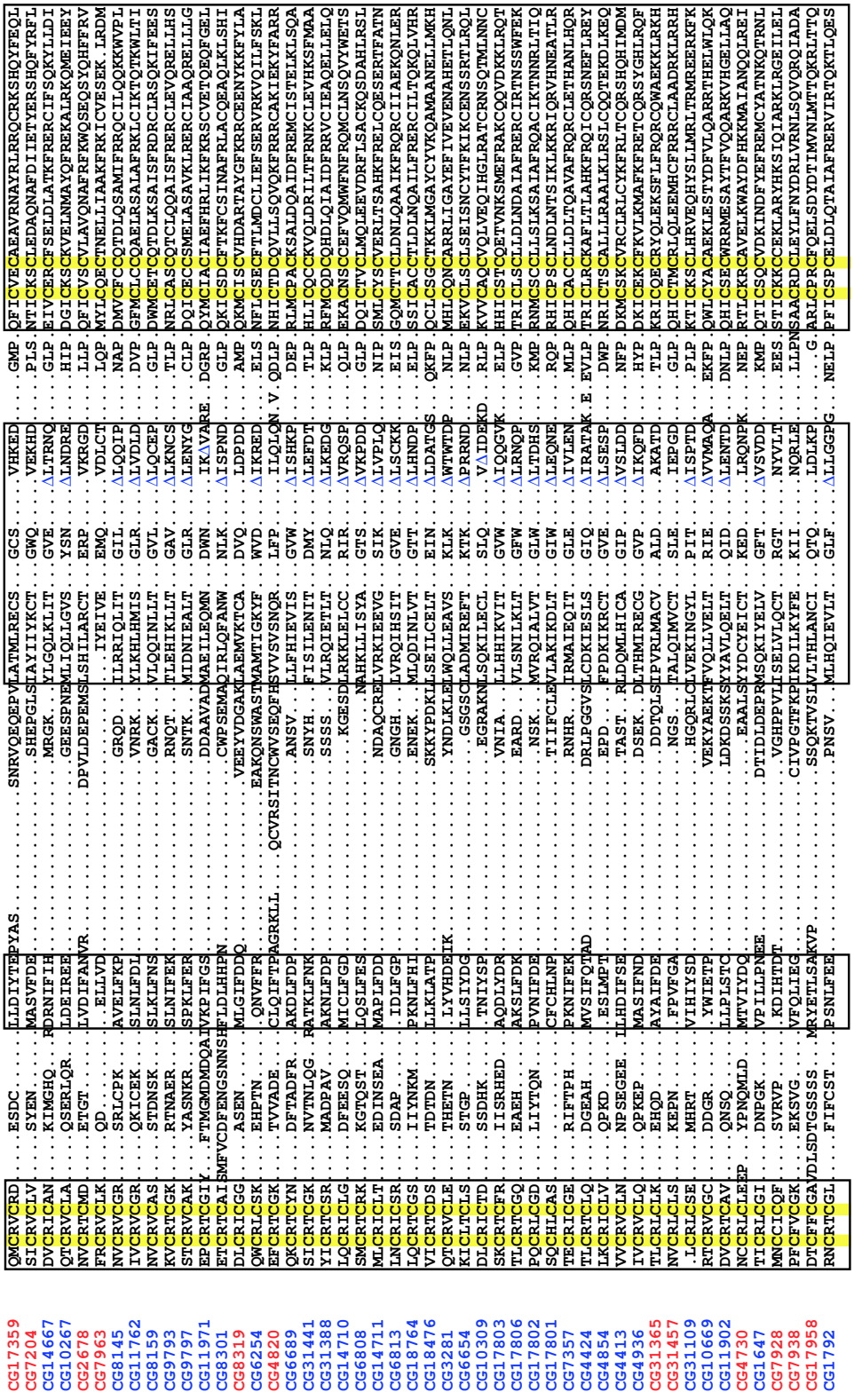


Figure B.1: Multiple sequence alignment of all *D. melanogaster* ZADs. Invariant cysteines are boxed in yellow; the conserved blocks are denoted by boxes; in red, subset 1 ZADs and in blue, subset 2 ZADs; the position of the intron is marked by a blue Δ.

Figure B.2 *continued on next page*

Figure B.2 *continued from previous page*

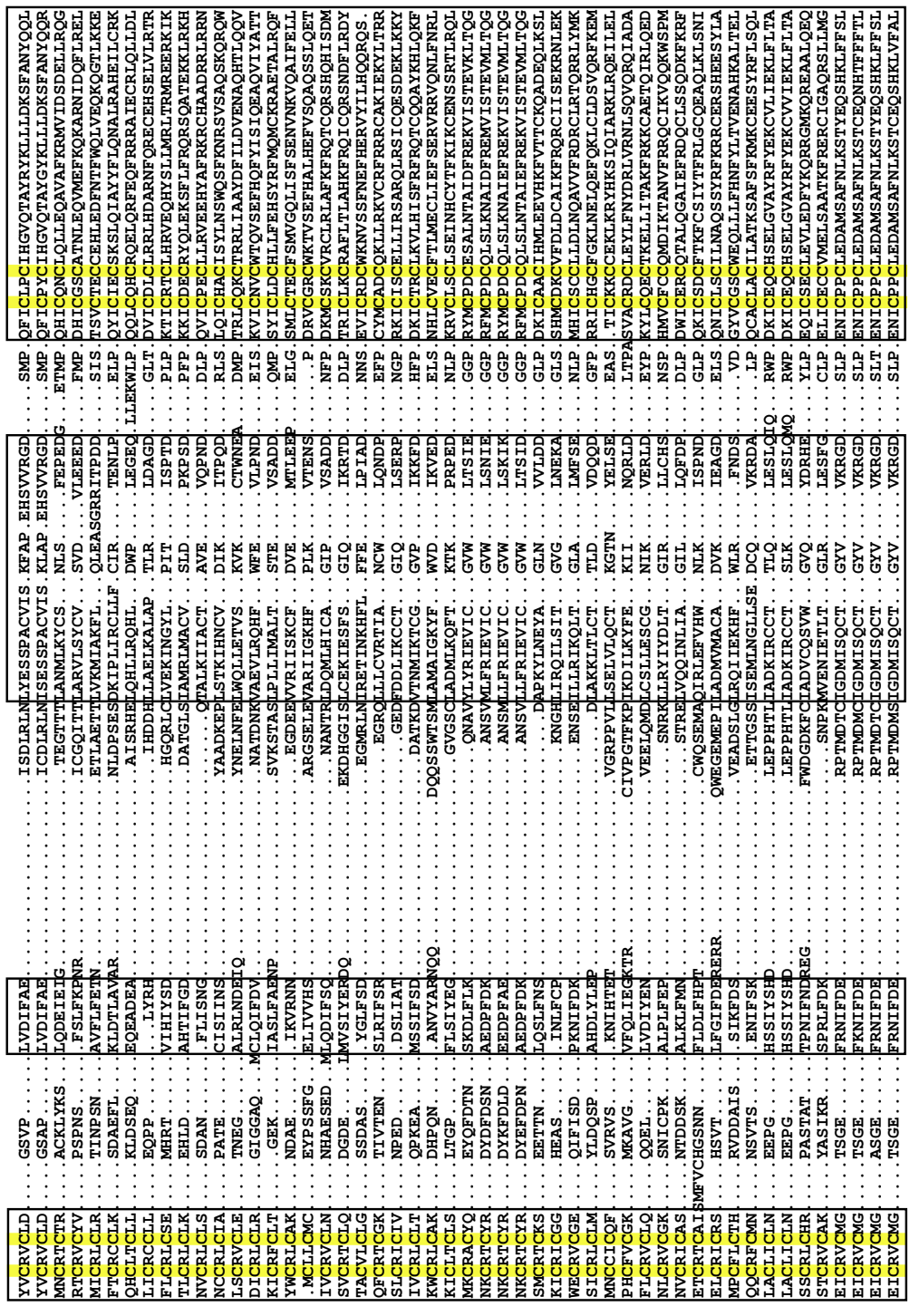


Figure B.2 *continued on next page*

Figure B.2 *continued from previous page*

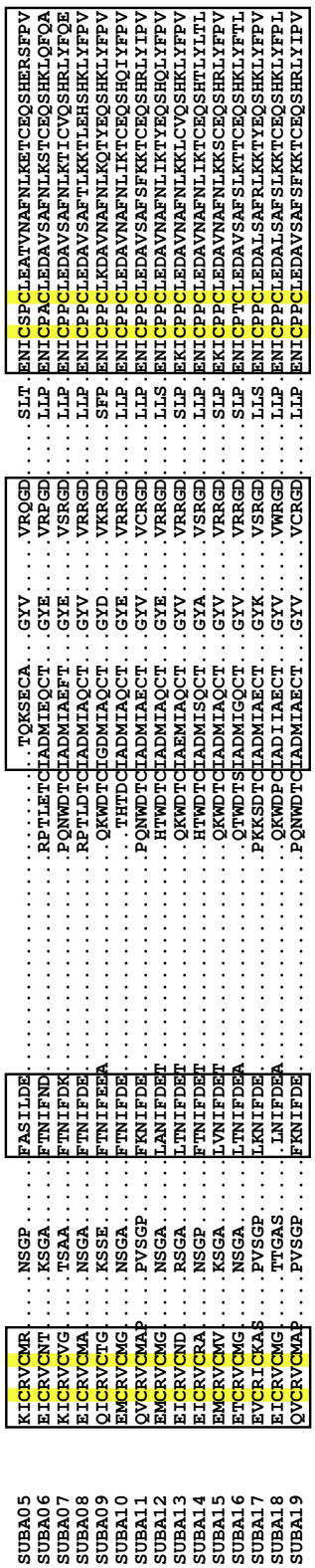


Figure B.2: Multiple sequence alignment of all *D. pseudoobscura* ZADs. Invariant cysteines are boxed in yellow; the conserved blocks are denoted by boxes.

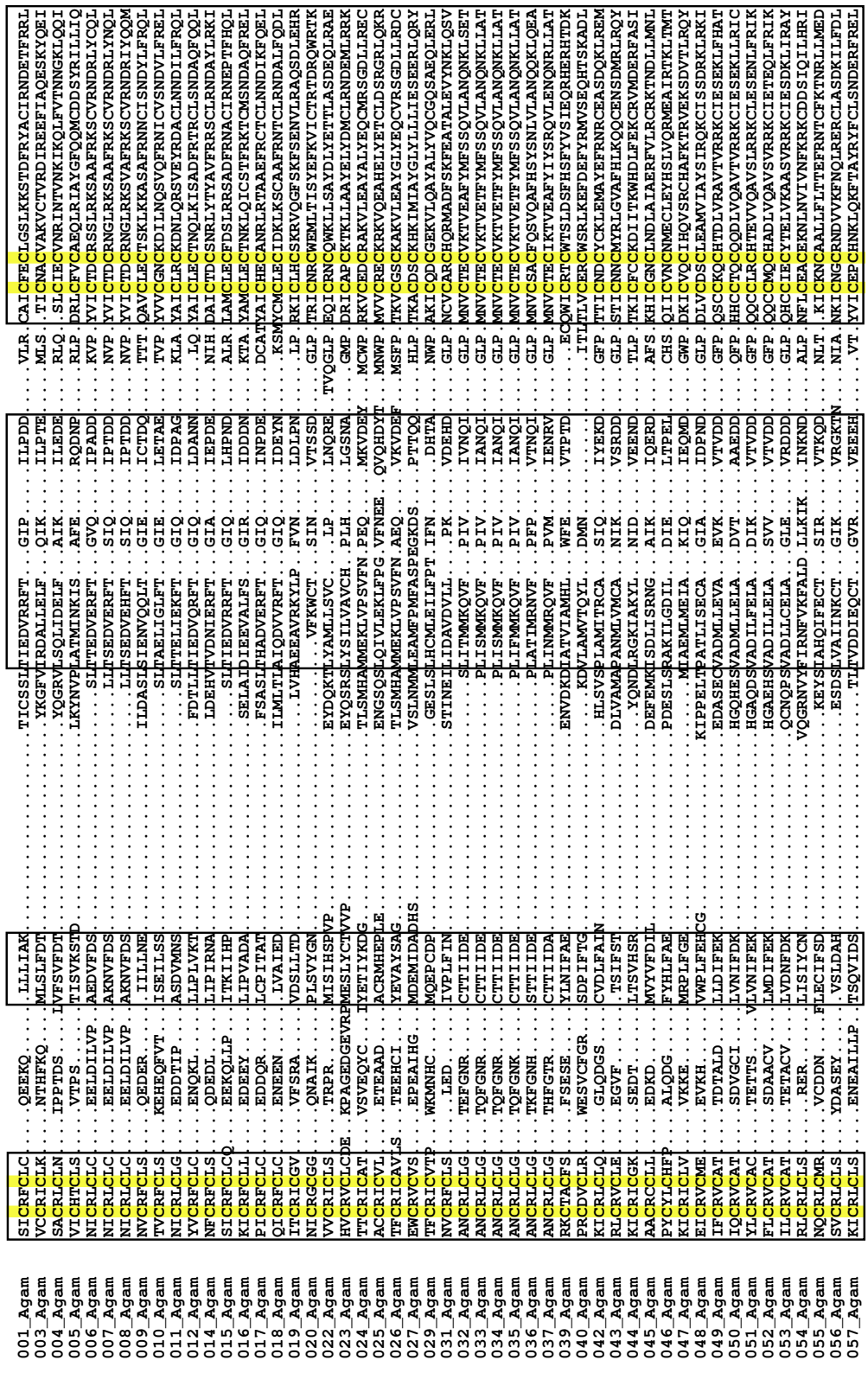


Figure B.3 *continued on next page*

Figure B.3 *continued from previous page*

Figure B.3 *continued on next page*

Figure B.3 *continued from previous page*

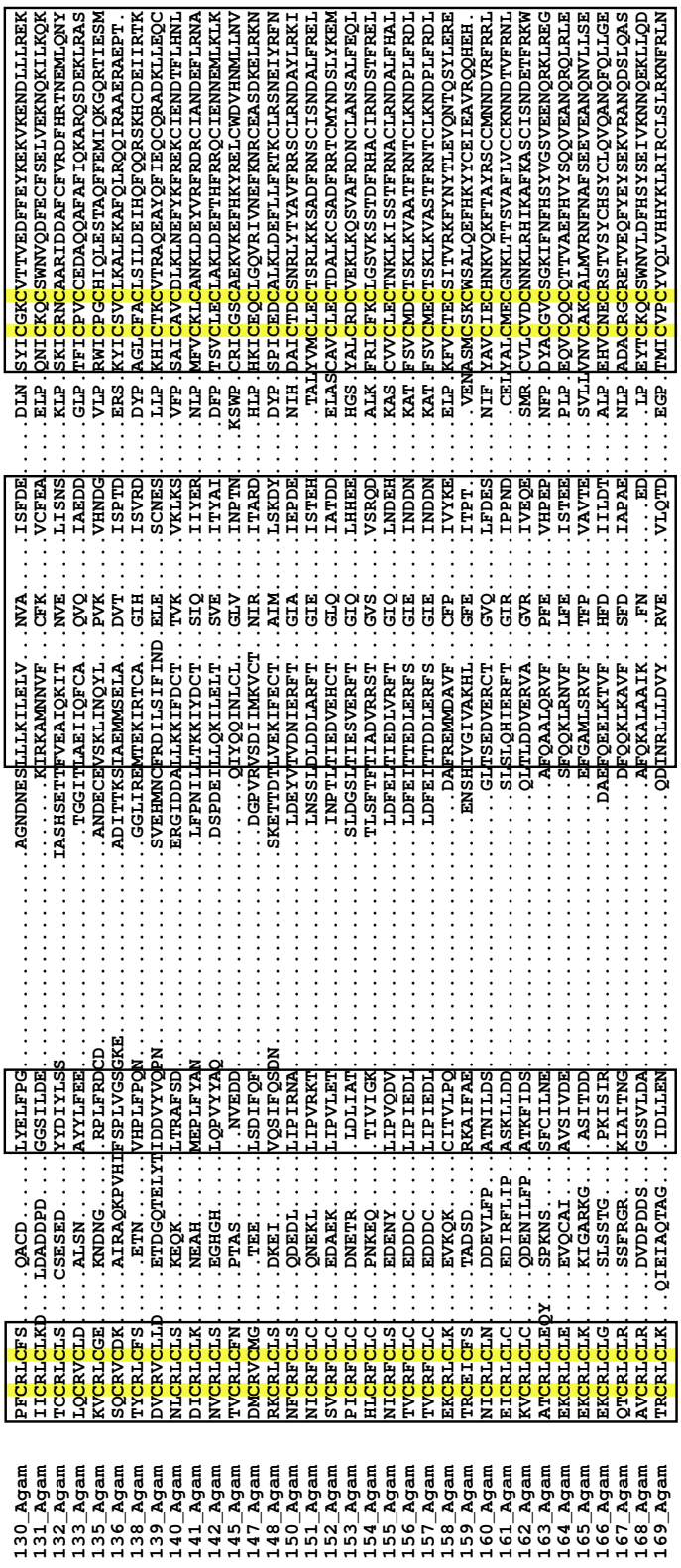


Figure B.3: Multiple sequence alignment of all *A. gambiae* ZADs. Invariant cysteines are boxed in yellow; the conserved blocks are denoted by boxes.

B.2 Neighbour-joining trees

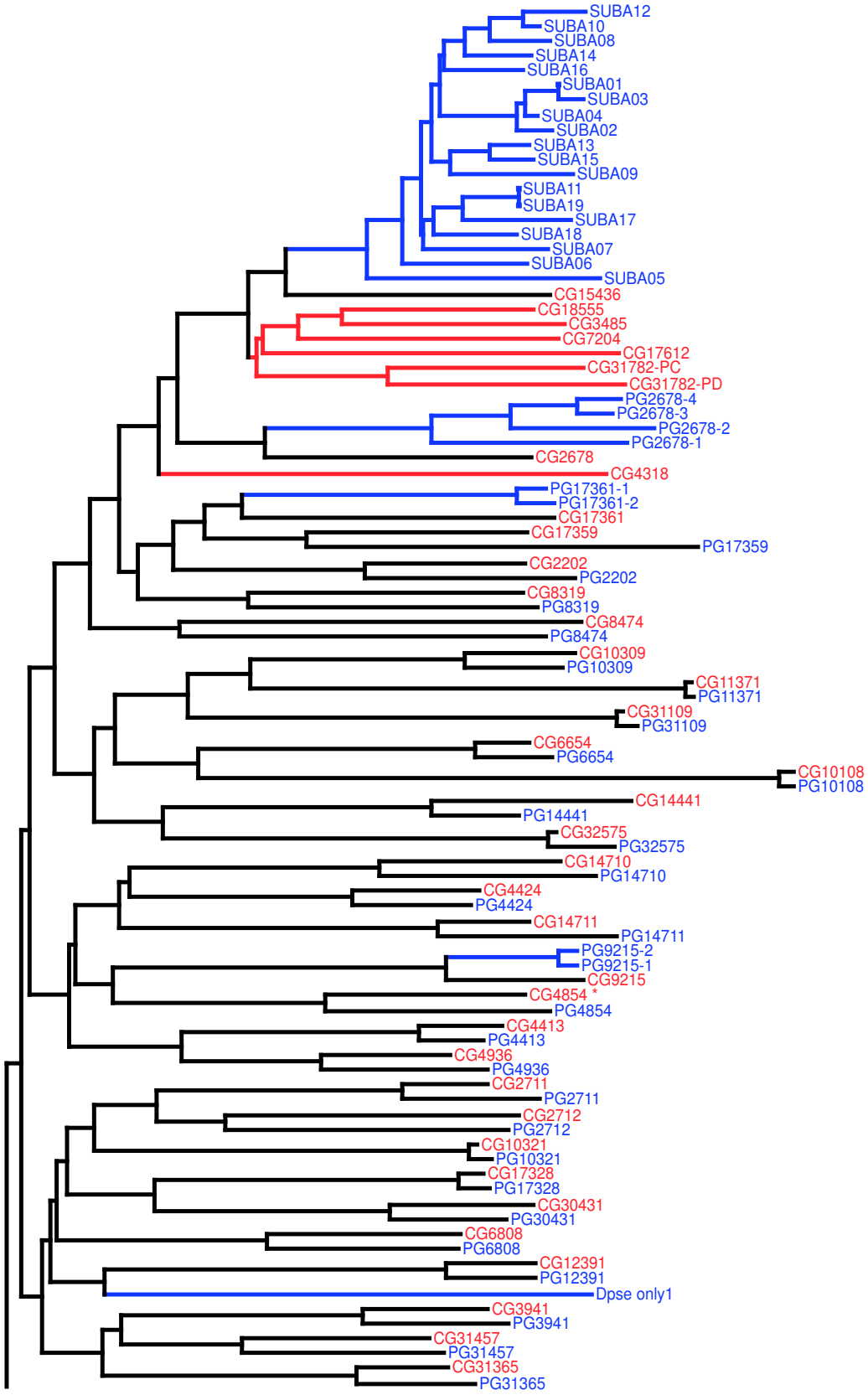


Figure B.4 *continued on next page*

Figure B.4 *continued from previous page*

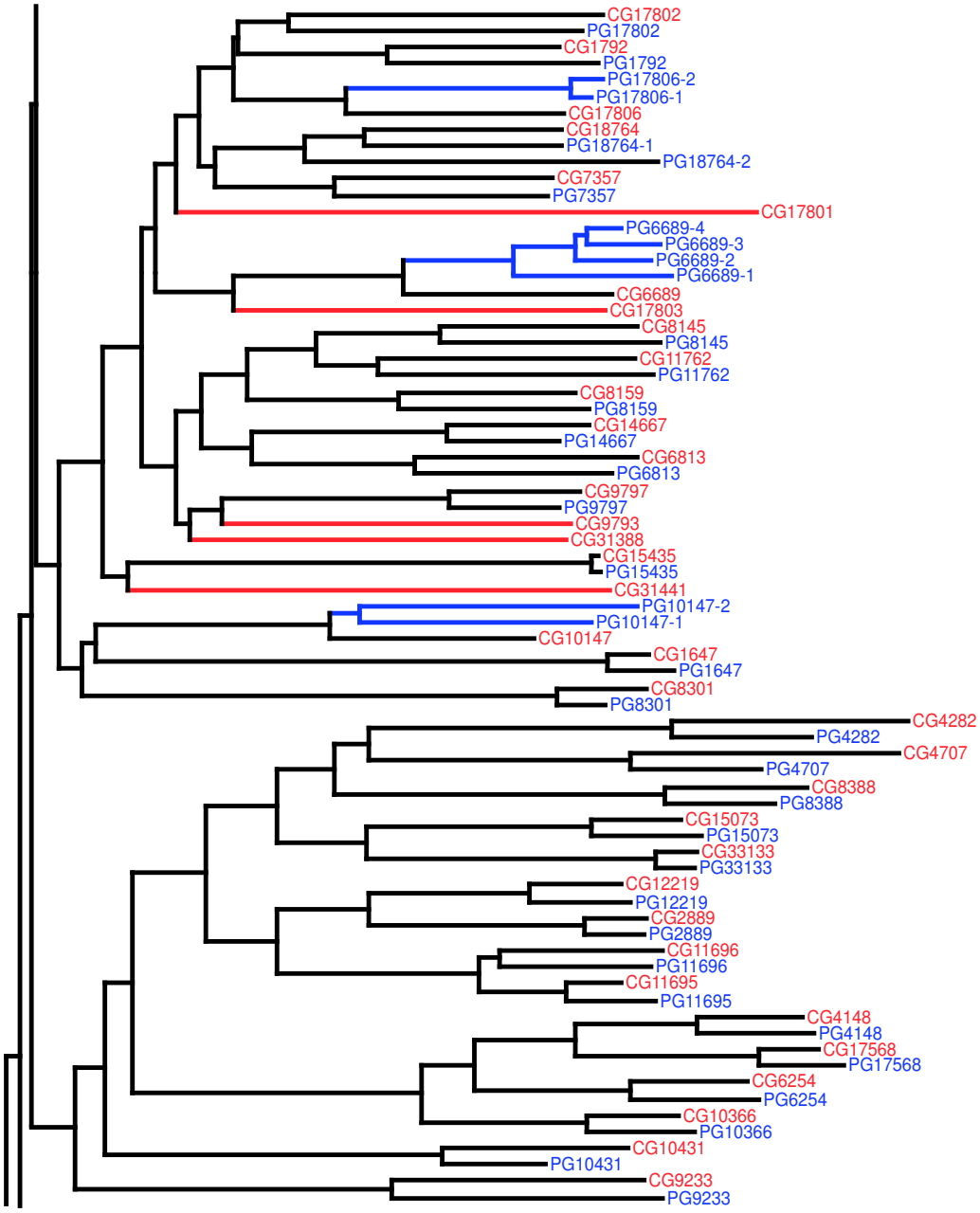


Figure B.4 *continued on next page*

Figure B.4 *continued from previous page*

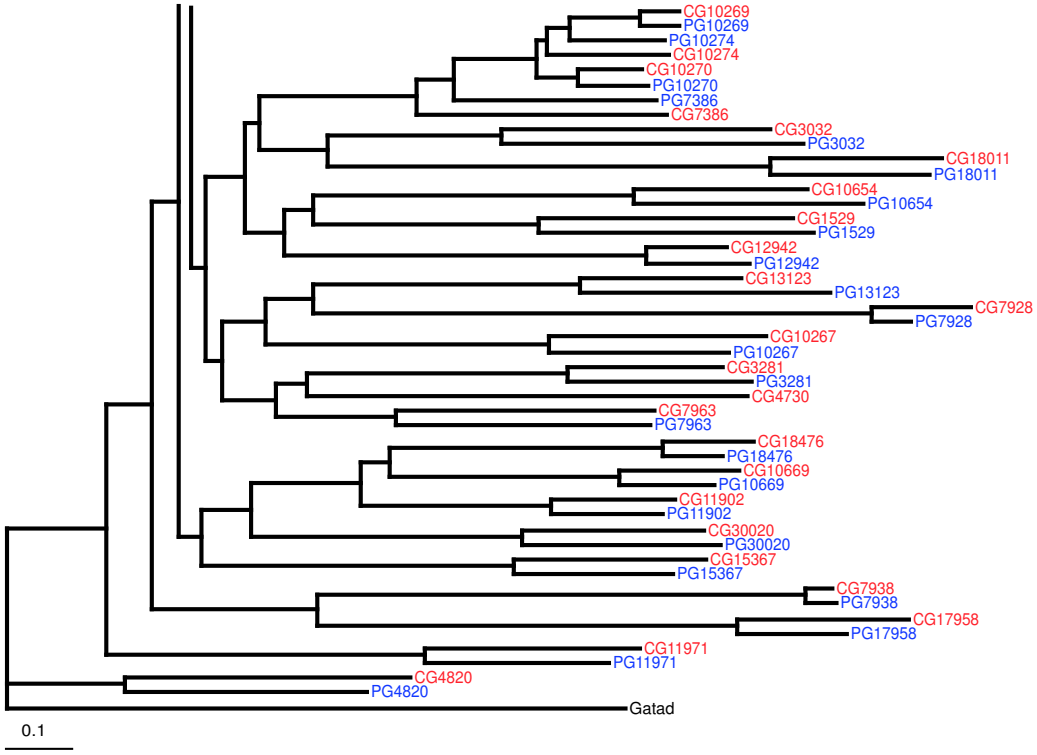


Figure B.4: *D. melanogaster* and *D. pseudoobscura* ZADs, full NJ tree. In red, *D. melanogaster* ZADs and in blue, *D. pseudoobscura* ZADs; red branches indicate *D. melanogaster*-specific branches; blue branches indicate *D. pseudoobscura*-specific branches; CG4854 is marked with an asterisk to indicate its migration from the subgroup d cluster to an alternative position; Gatad denotes a Gata zinc finger used to root the NJ tree; scalebar indicates the number amino acid substitutions per site.

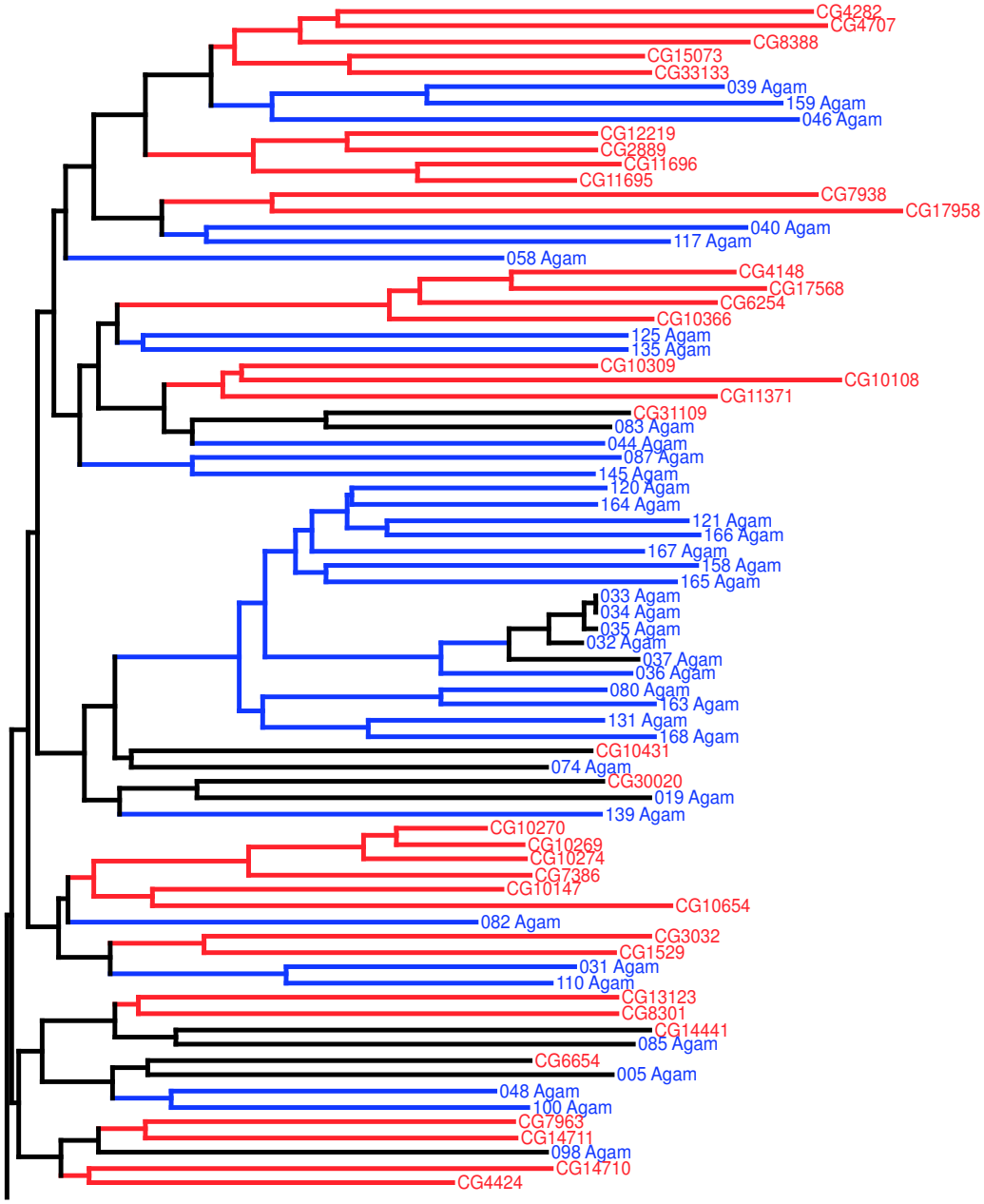


Figure B.5 *continued on next page*

Figure B.5 *continued from previous page*

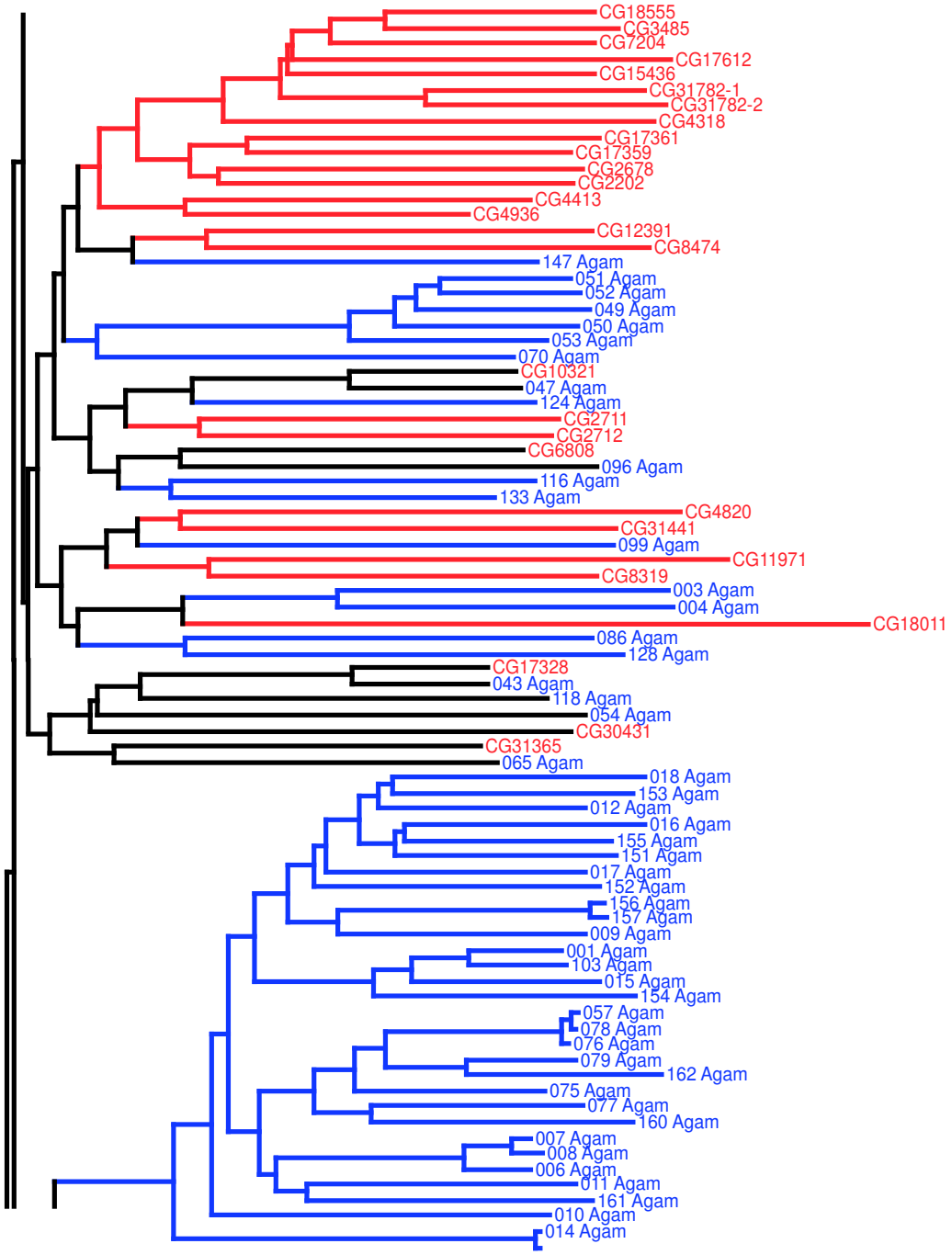


Figure B.5 *continued on next page*

Figure B.5 *continued from previous page*

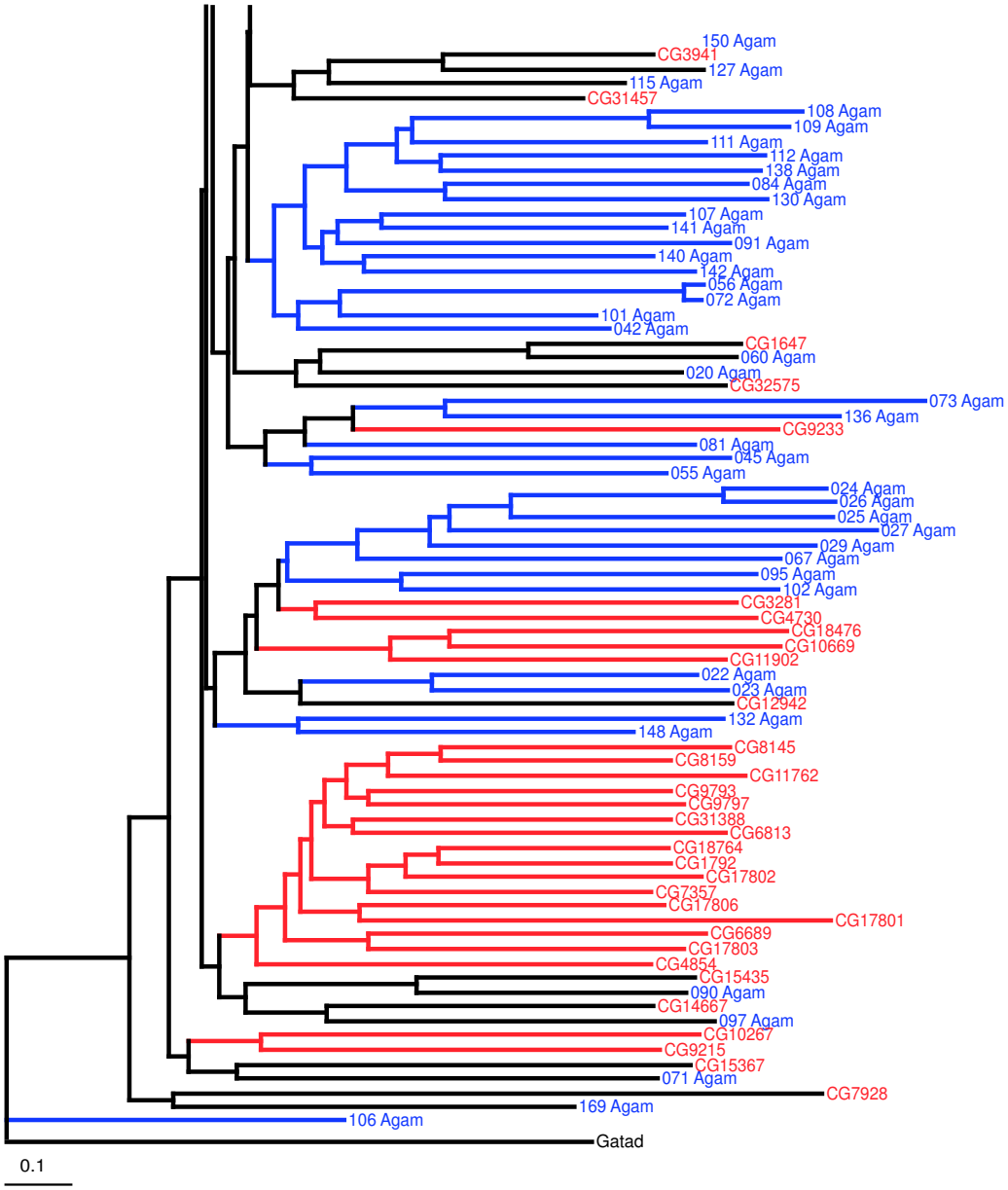


Figure B.5: *D. melanogaster* and *A. gambiae* ZADs, full NJ tree. In red, *D. melanogaster* ZADs and in blue, *A. gambiae* ZADs; red branches indicate *D. melanogaster*-specific branches; blue branches indicate *A. gambiae*-specific branches; Gatad denotes a Gata zinc finger used to root the NJ tree; scalebar indicates the number of amino acid substitutions per site.

Danksagung

Mein besonderer Dank geht an Herrn Prof. Dr. Herbert Jäckle, der mir diese Arbeit ermöglichte und durch lebhafte Diskussionen belebte.

Dem Boehringer Ingelheim Fonds möchte ich für die finanzielle Unterstützung während des Großteils meiner Promotion danken.

Herrn Prof. Dr. Dieter Jahn danke ich für die Vertretung der Arbeit an der Gemeinsamen Naturwissenschaftlichen Fakultät der Technischen Universität zu Braunschweig.

Ein ganz besonderer Dank geht an Dr. Siegfried Böhm für die schöne und gute Zusammenarbeit am ZAD-Projekt

Ralf Jauch danke ich für die Zusammenarbeit an der Bestimmung der räumlichen Struktur der ZAD des Transkriptionsfaktors Grauzone

Mitsuko danke ich für die Zusammenarbeit am SelB-Projekt.

Alexey danke ich für die Zusammenarbeit am *Krüppel* chromatin IP-Projekt

Weiterhin möchte ich mich bei allen Mitgliedern der Abteilung für Molekulare Entwicklungsbiologie für das angenehme Arbeitsklima bedanken. Insbesondere danke ich Ulrike Löhr, Ralf Pflanz und Ulrich Schäfer für die kritische Durchsicht dieser Arbeit.

Mein letzter Dank gilt meiner Frau Cornelia, die mich mit viel Gedult und Verständnis durch die letzten vier Jahre begleitet hat.